Developing omics and cell culture resources in non-model organism for their application in elucidating complex trait evolution in natural populations.

by

Amanda Denise Clark

A dissertation submitted to the Graduate Faculty of
Auburn University in
partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 10, 2022

Keywords: Genomics, Transcriptomics, Primary Culture, Comparative Biology, Biodiversity

Approved by

Tonia S. Schwartz, Chair, Associate Professor, Biological Sciences
Jamie Oaks, Associate Professor, Biological Sciences
Elizabeth Schwartz, Associate Professor, Biological Sciences
Amanda Sparkman, Associate Professor, Biology

Abstract

In the past, researchers chose models organisms to answer research questions based on their simplicity in morphology, domestication, and/or life history traits. Now, with high throughput sequencing (HTS) rapidly becoming cheaper, more investigations are tractable even with relatively small budgets. This is important because there are many complex and system-specific questions about natural populations and natural phenomenon that cannot be answered with traditional model organisms. Further, we cannot refine our understanding of life and biodiversity in a small unrepresentative subset of model organisms. The goal of this work is the generation of omics and cell culture resources for research on *Daphnia* and *Anolis* genera, respectively. I discuss their applicability for investigations of natural populations in the context of conservation and organisms themselves. Chapters One & Two develop omics resources in *Daphnia* systems, which are highly tractable evolutionary and ecotoxicology models. Chapter One, a published manuscript in G3, describes the generation of reference-guided draft genome assemblies for two strains of *D. pulicaria* with differential responses to toxic algal blooms. These assemblies with a high-quality curation of genes can be used for several downstream investigations, including the exploration of differential gene expression in response to algal toxins and identification of sequence variants associated with toxin resistance. Chapter Two diverts focus to the technical side of omics investigations where I explore 18 different combinations of RNA-seq tools for DGE analysis using a computationally tractable caloric restriction data set from *D. pulex*. I discuss the variation in biological or functional results due to the different tool combinations and explore variation at each step of the pipelines. The work presented in this chapter is the basis for three manuscripts in development: (1) the current chapter to be expanded and contrasted across species; (2) the biological inferences from the

RNA-seq results in conjunction with phenotypic data; and (3) the code used for these comparisons as a detailed tutorial for instructors to teach these analyses using a non-model system. My final chapter details the generation of cellular resources for testing *in silico* or omics generated functional predictions. I develop primary and early passage cells for > 100 lizards from the *Anolis* genus, providing methods for establishment and validation of reptile primary cells. I discuss their applicability in the context of studying protected and/or cryptic species in a dish!

Acknowledgments

I have had a host of individuals supporting me from the day I was born. Each one of them has made a significant impact on my life that nudged me down paths that ultimately led me to this point. I would like to highlight my family, as they have always prioritized my education – specifically one I am passionate about. I'd like to dedicate this work specifically to my late father, as I'm sure he would be thrilled to see where I am today.

I would like to acknowledge Verna Gates, Executive Director of Fresh Air Family. She shares my passion for education and the natural world and was a significant supporter of my graduate career from start to finish.

I would like to acknowledge Dr. Tonia Schwartz and Dr. Daniel Warner for allowing me the space to delve into a field unbeknownst by me, that would change my life forever! Tonia is an amazing person and advisor that has pushed me to do things I would never have the courage to. I never felt like I lacked support or opportunity, so I thank you Tonia.

I would like to acknowledge my wonderful committee Dr. Amanda Sparkman, Dr. Jamie Oaks, and Dr. Elizabeth Schwartz. These individuals have provided me with an amazing support system and much-needed reality checks. I had the most wonderful opportunities for field work that allowed me to spend time with Amanda S. I consider her family an extension of my own.

I would like to acknowledge Dr. Bruce Smith, my outside reader. Bruce's willingness to critique my work and his share excitement for my research boosted my confidence in preparation for my defense.

I would like to acknowledge Dr. Melody Russell who has been cheering me on my entire time at Auburn and provided me with professional development opportunities through NSF AGEP.

I would like to acknowledge Dr. Laurie Stevison who has been like a secondary advisor, training me in the skills I rely heavily on, and supporting and sharing my passion for education.

I have many friends, mentees & associates that have made my time at Auburn memorable and I would like to thank each one – but that is a long list of individuals. I would like to thank my lab mates and now close friends for rallying behind me, even when I did not do it for myself.

I would like to acknowledge Auburn University and the National Science Foundation for two fellowships (CMB and GRFP, respectively) that allowed me to focus on the many facets of my graduate career and general interest.

Lastly, I would like to acknowledge the Auburn High Performance Computing staff and resource. They have tolerated (happily) me stumbling through HPC best practices and have provided support in any way they could.

Thank you to everyone I have had the opportunity to come across on my journey.

Table of Contents

Abstract 3

Acknowledgments..... 5

List of Tables 10

List of Figures 11

List of Abbreviations 13

Draft genomes for one *Microcystis*-resistant and one *Microcystis*-sensitive strain of the water flea, *Daphnia pulicaria* 23

Introduction 24

Materials & Methods 26

Results & Discussion 32

Conclusions 37

References 41

Comparison of 18 common RNA-seq pipelines for differential gene expression analysis: from mapping to functional pathway enrichment 51

Introduction 52

Materials & Methods 56

Results & Discussion 65

Future Directions & Conclusions 84

References 89

Primary culture in non-model organisms: the establishment and validation of *Anolis* lizard dermal cells 108

Introduction 109

Materials & Methods 113

Results & Discussion 120

Future Directions & Conclusions 126

References 129

Conclusions 137

References 145

Appendix 1: Supplementary Material 153

Appendix 2: Abstracts of Additional Contributions..... 191

List of Tables

Table 1.1. A table of statistics from reference-guided assemblies for *Daphnia pulicaria* 49

Table 2.1: Pipeline Programs – Descriptions and Parameters Used 97

Table 2.2: Analysis of Variance (ANOVA) summaries of model 1 for each filtering method 104

Table 2.3: Linear model 1 summaries for each filtering method 105

Table S2.1 Contrast between pipelines from emmeans for Model 1 155

Table S2.2: Analysis of Variance (ANOVA) summaries of model 2 for each filtering method 173

Table S2.3: Linear model 2 summaries for each filtering method 174

Table S2.4 Contrast between pipelines from emmeans for Model 2 175

Table S3.1: Details of provenance and demographics of Anoles established using these methods 187

List of Figures

Figure 1.1. Percent of PA42 Genes for Different Average Depths of Coverage 33

Figure 1.2. BUSCO Analysis for D. pulicaria Genome Assemblies Indicate High Levels of Gene Content in Draft Assemblies 35

Figure 1.3: BlobPlots Indicate Low Levels of Contaminant Phyla in BA411 and WI6 Draft Reference-Guided Assemblies..... 36

Figure 2.1: RNA-seq Pipelines for Differential Gene Expression Analysis Compared 65

Figure 2.2: Mapping Statistics for Hisat2 and STAR aligners across samples 67

Figure 2.3: Raw count distributions (diagonal) and correlations across quantification methods for 2 samples 69

Figure 2.4: High similarity between quantification methods across samples based on Rv coefficients 71

Figure 2.5. Plotted linear model 1 summaries from Table 2.3 73

Figure 2.6: HTSeq and Limma-Voom increase the number of significant DEGs detected across pre-filtering methods 75

Figure 2.7. Plotted linear model 2 summaries from Table S2.3 77

Figure 2.8. High overlap between DGE programs, particularly those using Negative Binomial Models (Pipeline) 79

Figure 3.1: Anole cells at different stages of establishment and culture 119

Figure 3.2: Example of Short Tandem Repeat (STR) profiles generated from cell and tail DNA from the same individual match 121

Figure 3.3: Anole primary cells have population doubling times of ~2.4 days 123

Figure S1.1: FastQC Screen Analysis Indicates the Expected Composition of Reads Based on Screened Genomes 153

Figure S1.2: Sourmash Distance Estimates Indicate Higher Similarity Between D. pulicaria Strains (BA411 & WI6), Relative to the D. pulex (PA42) Reference 154

Figure S2.2. High overlap between DGE programs, particularly those using Negative Binomial Models (Soft) 185

Figure S2.3. High overlap between DGE programs, particularly those using Negative Binomial Models (Hard) 186

List of Abbreviations

HTS	High-Throughput Sequencing
GH-IIS	Growth Hormone and Insulin/Insulin-Like Signaling
IGF	Insulin-Like Growth Factor
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
DEG	Differentially Expressed Genes
DGE	Differential Gene Expression
GSEA	Gene Set Enrichment Analysis
PA42	*Daphnia Pulex* Reference Genome
GTF	Gene Transfer Format File
FBS	Fetal Bovine Serum
CS	Chicken Serum
IPSC	Induced Pluripotent Stem Cell
UVA	University of Virginia Arlington
DMSO	Dimethyl sulfoxide

INTRODUCTION

1.1 Motivations to Study Phenotypic Variation

The desire to understand the complexities of life have beckoned interest and efforts of researchers across the biological sciences before the time of Rosalind Franklin, Robert Hooke, or even Aristotle (Voultsiadou *et al.* 2017; Vucetich *et al.* 2021). With the contributions of these few individuals and the countless more that came before and after their time, we have learned that our world consists of a mind-blowing amount of variation and diversity at every level of the biological hierarchy, from gene to cell to organism phenotypes. We have explored many questions at these biological levels, whether those questions addressed the breadth of phenotypic variation on Earth (Gerovasileiou *et al.* 2015; Anderson 2018), how new phenotypes arise and disperse through populations (Orr 1998; Fox *et al.* 2019; Card *et al.* 2019), or the mechanisms (i.e., ecological and/or molecular) underlying phenotypes (Christe *et al.* 2000; Schlichting and Smith 2002; Lomolino 2005; Schwander *et al.* 2014; Funk *et al.* 2016; Herrera-Álvarez *et al.* 2018; Benítez-López *et al.* 2021). Often these investigations, along with the progression of technology, have generated just as many new questions as they have answers! While there is still a plethora of questions in individual biological fields, what has fascinated me most are the questions that require integrative approaches across fields and biological hierarchy. These tantalizing questions and investigations can tell the intricate stories of life, representing both its connectivity and diversity.

Even with the incredible amount of variation in life, there are still many conserved or convergent biological systems across numerous, distinct groups of organisms. As an example, consider metabolic regulation in traditional model systems (yeast, fly, mouse, nematode) and other systems (cow, chicken, dogs, snakes). There is an overwhelming amount of evidence for the GH-IIS pathway as a conserved molecular system regulating this trait across these very different taxa (Barbieri *et al.* 2003; Metallo and Heiden 2013; Khan *et al.* 2014; Perry *et al.* 2019; Fujita *et al.* 2019; Chowański *et al.* 2021; Chatterjee and Perrimon 2021). Further, the GH-IIS regulates, in part, the confounded traits of growth, reproduction, and aging at the level of the cell and the organism (Barbieri *et al.* 2003; Denley *et al.* 2005). For me, contemplating these ideas always brings me to the question "how does such extensive phenotypic diversity arise from seemingly constrained genetic sequences and molecular networks?"

Understanding how evolutionary forces, like natural selection, acts across molecular networks may provide more comprehensive insight on the evolution of complex traits, or traits regulated and determined by multiple genes and environmental interactions, that are often regulated by these networks. The field has benefited from the many molecular evolution studies to date that endeavor to provide this insight, even though they mostly focus on a single to a few candidate genes (Hoekstra 2006; Hsu *et al.* 2008; Radwan and Babik 2012). While all efforts are integral to our understanding of how complex phenotypic traits evolve, the omission of environmental and multigenic interactions limits our detection of factors that contribute to the evolution of the novel phenotypes vital to Earth's biodiversity. The relatively minor lag in omnigenic approaches to molecular

evolution studies in well-funded, established model organisms could be attributed to our limited ability to mathematically model these networks and the tractability of studying mendelian traits (Glazier *et al.* 2002; Belmont and Leal 2005; Mathieson 2021). However, in systems that deviate from the traditional models, this issue is compounded by a lack of resources to address questions in these less-developed systems.

1.2 Motivations to Use Non-model Organisms

Biological research is seemingly dominated by a few study systems, deemed model organisms. 'Model organism" was a term originally reserved for taxa that had certain levels of tractability for scientific investigations (Müller and Grossniklaus 2010). Typically, this tractability was in the context of an organism's reproductive rates, size, ease of genetic manipulation, and ease of obtainment or handling or culturing in laboratory settings. Today, "model organism" is used ubiquitously for taxa supported by extensive system knowledge and/or resources (standard methodology, databases, informatics) and they often address a specific biological process or phenomenon (Ankeny and Leonelli 2011). Model organisms provide an invaluable platform for scientific investigations. We frequently solve problems using reductionist methods, with the idea that removing complexities can provide more direct investigation of the problem at hand (i.e., modelling). This was one motivation of developing the bacteria *Escherichia coli* as a model organism. Prior to the 1950s emergence of molecular biology, *E. coli* were already being extensively used in viral and microbiological investigations. Following the emergence of molecular biology came foundational knowledge of DNA replication, gene

expression, and restriction enzymes, all initially generated in *E. coli* models (Blount 2015). Many questions pertaining to eukaryotes, let alone vertebrates, could not be answered in this prokaryotic model. This required increasingly complex, diverse models like yeast, flies, and mice. These models have clarified and resolved many basic molecular biology questions about eukaryotes and have been extensively used as disease models in the biomedical sciences (Veldman and Lin 2008; Howe *et al.* 2013; Menezes *et al.* 2015; Blount 2015; Baldridge *et al.* 2021). While I cannot deny the indispensability of model organisms, we still cannot ignore how much they have contributed to investigative tunnel vision. Due to their tractability, more studies were performed in model organisms and concomitantly more resources were developed and optimized for these limited number of systems. This eventually created a culture of using these models to answer questions that they may not have been suitable for because there were adequate resources for them and an assumption that model organisms would be broadly representative across taxa (Leonelli and Ankeny 2013; Seifirad and Haghpanah 2019). Developing a model is not trivial, both in the context of effort and costs when you consider the need for characterization of an organism (i.e., anatomically, ecologically), molecular tools compatible with an organism (i.e., antibodies, biomaterials, standardized protocols), breeding/husbandry, and information infrastructure (i.e., databases, biological products, and molecular data). Yet, we are limited in the depth and scope of our investigations of biodiversity without adequate access to numerous, well-distributed representatives across the Tree of Life. In science, we base our understanding on the outcomes of observation

and experimentation, but how complete can our understanding of life actually be when we have a been focusing on a narrow collection of organisms?

1.3 How Technology is Bringing Balance

We are already seeing researchers step out of their comfort model organisms, developing new systems to answer questions about unique biological phenomena. This is largely due to the advent of massively paralleled or high throughput sequencing (HTS) technologies that has allowed large scale investigations of genes, proteins, metabolites, and other molecules of interest (Soon *et al.* 2013; Reuter *et al.* 2015). Advances in biotechnology and biochemistry have driven the cost of HTS lower and lower, opening the door for non-model investigations in the omics era (Ellegren 2014; Reuter *et al.* 2015; Muir *et al.* 2016). Omics refers to the global assessment of a particular set of molecules (i.e., genomics – DNA, transcriptomes – RNA, proteomics – Protein) in an organism (or tissue or single-cell) at the time of sampling. Personally, it is an exciting time to be an evolutionary biologist with the potential to apply comparative omics and cell biology to explore the mechanisms and evolution of regeneration or apply metagenomic sequencing to continuously resolve basal relationships on Tree of Life (Spang *et al.* 2017; Stockdale *et al.* 2018). Of course, model organisms were the first up for access to these technologies due to the existing infrastructure, which arguably made it more feasible to direct funding to these systems. Still, sequencing in non-model organisms have advanced studies focused on biodiversity (Bonasio 2015; Székely 2019; Burnett *et al.* 2020), aided in the progress of population genetics from a largely theoretical to an empirical field (Ellegren

2014), and assisted in uncovering taxa better suited than existing model organisms to address more central, longstanding biological questions (Russell *et al.* 2017). With these successes comes new challenges of data analysis and testing these functional predictions and hypotheses experimentally. The access to developed molecular protocols, databases, and biomaterials is still a large hinderance in non-model systems, as is access to formalized computational biology, bioinformatic, and data management training necessary to analyze and handle modern biological data. This is compounded by an inability to escape our investigative tunnel vision, as several well-maintained bioinformatic programs, methods, and computations resources were developed for specific taxa (usually human) and are not well optimized for organisms with different genomic architectures (i.e., repetitive elements, ploidy).

My dissertation reflects my desire to appreciate the development and evolution of complex traits through an understanding of underlying genetic network interactions and molecular constraints contributing to phenotypic variation. Along my journey, I have realized the dearth of optimized protocols and resources available for the systems in which I chose to address my evolutionary questions. This has promoted a pivot in my work to resource development and, as a secondary theme of my research, promoting the development of new models to study cellular physiology and functional genomics in the context of ecology, evolution, and ultimately biodiversity.

1.4 Chapter Abstracts

In my first chapter, I generate genomic resources for the study of adaptive toxin resistance in water fleas! *Daphnia* species are well-suited for studying local adaptation and evolutionary responses to stress(ors) including those caused by algal blooms. Algal blooms, characterized by an overgrowth (bloom) of cyanobacteria, are detrimental to the health of aquatic and terrestrial members of freshwater ecosystems. Some strains of *Daphnia pulicaria* have demonstrated resistance to toxic algae and the ability to mitigate toxic algal blooms. Understanding the genetic mechanism associated with this toxin resistance requires adequate genomic resources. Using whole genome sequence data mapped to the *Daphnia pulex* reference genome (PA42), we present reference-guided draft assemblies from one tolerant and one sensitive strain of *D. pulicaria*, Wintergreen-6 (WI-6) and Bassett-411 (BA-411), respectively. Assessment of the draft assemblies reveal low contamination levels, and high levels (95%) of genic content. Reference scaffolds had coverage breadths of 98.9% - 99.4%, and average depths of 33X and 29X for BA-411 and WI-6, respectively. Within, we discuss caveats and suggestions for improving these draft assemblies. These genomic resources are presented with a goal of contributing to the resources necessary to understanding the genetic mechanisms and associations of toxic prey resistance observed in this species.

Chapter Two focuses on technical challenges of navigating the rapidly increasing number of computational and statistical tools used in transcriptomic and gene expression analyses. Here, I compare common tools used in the literature in different combinations to build 18 different RNAseq analysis pipelines. These types of studies have been previously done, but the majority make fewer comparisons or stop at differential gene

expression (DGE) analysis. I extend previous work to include functional gene set enrichment analysis (GSEA) from each of the pipeline to determine if there are changes in the biological interpretation of functional pathway enrichment results. The results confirm that while there is high similarity between mapping tools, there are significant effects of counting and DGE analysis methods on the number of differentially expressed genes (DEGs). Further, these differences carry over to functional pathway enrichment results, with the most noticeable effect of filtering choices on the ability to detect enriched gene sets at a reasonable FDR. Although there were limitations in enrichment detection due to annotation quality, several pipelines report enrichment of genes in the xenobiotic metabolism set. This finding was quite relevant to the caloric restriction data set used, as *Daphnia* metabolism and stress responses utilize the same receptors for both xenobiotic and endobiotic stimulants. These analyses have been performed in a reproducible manner, with tutorial-like scripts develop as a resource to learn how to perform DGE analysis across different common programs.

Finally, in Chapter Three I develop methodology to build cellular models from non-model organisms to circumvent conservation limitations in research, while adhering to the 3 tenets of animal research. Primary explant cell culture, growing cell monolayers from tissue sample, provides a method for studying cellular physiology and biochemical function. Primary cells have shorter replicative lifespans compared to immortalized cell lines, but they retain features programmed by the *in vivo* background, proving cell culture useful for testing cellular responses, and interpreting organismal responses to stimuli. Published methods for cell culture in ectothermic vertebrates are limited, and even fewer

are specific to non-avian reptiles. I detail method used for establishing over 100 different primary and early passage cells from opportunistic collections of lizard tails across multiple *Anolis* genera. I also detail methods for validating these cells prior to their use as experimental resources and discuss the avenues in which cellular resources could contribute to non-model investigations. Resources, including methodology, biomaterials, and expertise, are currently limiting factors for the broad inclusion of studies at the cellular level in evolutionary ecology research, making resources like the methods presented in this work critical to the field.

My desire is for this work to be contributory in advancing investigations in non-model systems to promote more integrative and comparative investigations of biodiversity and life itself.

CHAPTER I

Draft genomes for one *Microcystis*-resistant and one *Microcystis*-sensitive strain of the water flea, *Daphnia pulicaria*

Amanda D. Clark[*], Bailey K. Howell[†, ‡], Alan E. Wilson[§], Tonia S. Schwartz[*]

[*] Department of Biological Sciences, Auburn University, Auburn, AL 36849

[†] Bioinformatics REU program, Department of Biological Sciences, Auburn University, Auburn, AL 36849

[‡] Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24061

[§] Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL 36849

Published in G3:

Amanda D Clark, Bailey K Howell, Alan E Wilson, Tonia S Schwartz, Draft genomes for one *Microcystis*-resistant and one *Microcystis*-sensitive strain of the water flea, *Daphnia pulicaria*, *G3 Genes|Genomes|Genetics*, Volume 11, Issue 11, November 2021, jkab266, https://doi.org/10.1093/g3journal/jkab266

INTRODUCTION

Over the past two decades, functional ecology research has focused on constructing a theoretical framework for eco-evolutionary dynamics, the bidirectional feedback between ecological and evolutionary processes of populations, communities, and ecosystems (Pelletier *et al.* 2009). An integral portion of this framework requires connecting genetic variation in species and the concomitant effects on ecological interactions across hierarchal levels (Brunner *et al.* 2019). Ideal species for studying this interplay of evolution and ecology possess high connectivity within their ecological communities (see Figure 2 in Miner *et al.* 2012), experimental flexibility and tractability (i.e., responds to a multitude of diverse stressors; ease of controlled maintenance and manipulation), and suitable genomic resources (Miner *et al.* 2012). *Daphnia*, commonly known as water fleas, satisfy these criteria. *Daphnia* are well-studied, widely employed models in ecology, evolution, and ecotoxicology (Shaw *et al.* 2007; Eads *et al.* 2008; Sarnelle *et al.* 2010; Miner *et al.* 2012; Nelson *et al.* 2018; Asselman *et al.* 2018; Becker *et al.* 2018), and have been utilized in Nobel prize-worthy discoveries (Nobel Media AB 2021).

Daphnia species, with the appealing traits of short generation times and cyclic parthenogenesis, are well-suited for studying local adaptation and evolutionary responses to stress(ors) including those caused by global warming and anthropogenic eutrophication (Hairston *et al.* 2001; Ebert 2005; Asselman *et al.* 2014). *Daphnia pulicaria*, a lake-dwelling herbivorous zooplankton in the genus, demonstrate evidence of local adaptation

to cyanobacteria in eutrophic lakes and significant genetic structure amongst populations (Sarnelle and Wilson 2005; Chislock *et al.* 2019a). *Microcystis aeruginosa* is a highly toxic species of cyanobacteria, abundant in harmful algal blooms, that may produce toxic metabolites, including a suite of hepatotoxins called microcystins (Paerl *et al.* 2001). Many methods for controlling these blooms have been proposed due to their adverse effects on human health, the economy, and ecological communities. One promising avenue for mediation is biomanipulation or manipulating trophic levels to control cyanobacterial overgrowth by introducing *D. pulicaria* exhibiting resistance to toxic cyanobacteria (Wilson and Chislock 2013; Chislock *et al.* 2019b). These findings have contributed to a strong, growing interest in using *D. pulicaria* to understand the mechanistic link between genetic trait variation and ecological community dynamics, which could aid in informing mitigation tactics for harmful algal blooms. However, such efforts require increasing the available genomic resources.

Currently, there are full genomes assemblies for four *Daphnia* species*: D. pulex,* TCO (Colbourne *et al.* 2011a) and PA42 (Ye et al. 2017); *D. magna,* KIT (Lee et al. 2019) and XINB3 (Gilbert,D.G, unpublished [PRJNA298946]); *D. carinata*, WSL (Jia et al. 2020); and a *D. galeata* assembly [PRJEB42807] (Nickel *et al.* 2021). Published whole genome amplifications of single and pooled *D. pulicaria* adult and ephippia have been mapped to TCO, but the genomic resources presented here are the first genome assemblies for *D. pulicaria* assembled using the new and improved, PA42 reference genome (Lack *et al.* 2018). Here, we present two reference-guided assemblies from two

strains of *D. pulicaria*, one *Microcystis*-resistant strain, Wintergreen-6 (WI-6), and one *Microcystis*-sensitive strain, Bassett-411 (BA-411).

MATERIALS & METHODS

Samples

Two strains of *D. pulicaria*, WI-6 and BA-411, were initiated from a single individual isolated from small glacial lakes in southern Michigan during autumn 2004 and spring 2009, respectively (Chislock *et al.* 2019a, 2019b). Tolerant phenotypes were established by exposing neonates to toxic cyanobacterial diets where strains from highly eutrophic lakes, like WI-6, demonstrated reduced negative impacts on growth rates (Sarnelle and Wilson 2005). These strains have been maintained in clonal cultures in the freshwater ecology laboratory of Dr. Alan Wilson (AU) since isolation and were received from the Wilson Lab in December of 2018 for genome sequencing. For each strain, approximately 10-15 individuals were cultured in autoclaved 50 mL flasks loosely capped with foam stoppers and transferred to fresh food and water on a biweekly basis. Clonal populations for each strain were cultured at (23^{o}C) room temperature in (autoclaved) water from a nearby, oligotrophic reservoir (Lake Martin, AL), and fed a nutritious alga, *Ankistrodesmus falcatus*, *ad libitum*. As populations reproduced, offspring were quantified and separated into new flasks on a weekly basis. Offspring were allowed to mature before being transferred into diethylpyrocarbonate (DEPC) treated water

[VWR, USA] with no food for two days in order to clear their guts. Post starvation, 20 adults were pooled into 1.5 mL tubes in 1mL of a 1.5x DNA/RNA Shield [Zymogen, USA] and stored at 4^{o}C. For each strain, we used three tubes of 20 adult *Daphnia* for DNA extractions for genome sequencing.

DNA extraction

DNA was extracted within 24 – 48 hours of DNA/RNA Shield storage using the QIAamp UCP DNA Micro kit [QIAGEN,Germany] per manual instructions, with some modifications. Briefly, DNA/RNA Shield was removed, 10mL of kit proteinase k (half the recommended amount) and two 2.0mm silicate beads were added before samples were homogenized on a TissueLyser II [QIAGEN] for 1 minute at a frequency of 30 cycles(s^{-1}). The remaining steps of the manufacture's protocol were followed with DNA being eluted from the filter in a 20 mL volume. Independent DNA extractions from were performed over three weeks during March 2019 were frozen at -20^{o}C. For each strain, the DNA from three samples were pooled and concentrated in preparation for genome sequencing, thus the genomic sequence represents approximately 60 individuals, that are presumed to be clonal. DNA was quantification using the Qubit dsDNA High Sensitivity Assay kit [Thermo Fisher, USA].

Validation of Strain via Genotyping PCR

To validate that DNA samples were of the correct and single strain of origin, samples were PCR amplified for DP496, a microsatellite locus previously identified in Colbourne et al. (2004) and demonstrated to have discriminating allelic patterns between these two *D. pulicaria* strains (Wilson and Hay 2007; Chislock *et al.* 2019a). Primer sequences were obtained from the *Daphnia* Genomics Consortium, wfleabase (http://wfleabase.org/genomics/microsatellite/) (Colbourne *et al.* 2004, 2005). PCR reactions were carried out in 10mL volumes using 5mL of 2X GoTaq Green PCR Master Mix [Promega, USA], 0.3 mL of 10mM forward and reverse primers (0.3 mM final concentration), 3.65mL of water, and 0.75 mL of DNA (21 ng). The thermocycler program for the PCR began with a 2 min denaturation cycle at 95ºC, followed by 35 cycles of 20 sec at 95ºC, 20 sec at 50ºC, and 20 sec at 72ºC, and a final extension cycle for 10 min at 72ºC. DNA from a single individual of each strain was used as positive controls, and water instead of DNA was used as a no template control. Five mL of PCR products were visualized on a 3% agarose gel made with 0.5X TAE and 1mL of GelGreen Nucleic Acid Stain [Biotium, USA] to confirm that the allelic patterns for the genome samples were consistent with the positive controls for the target strains.

Sequencing

For each strain, approximately 0.8 mg (WI-6) and 1 mg (BA-411) of DNA were shipped to Novogene [China] for sequencing. Novogene performed library preparation

using the Illumina TruSeqLibrary Construction Kit and sequencing on an Illumina Novoseq 6000, producing 8 Gbs (54.8 and 56.1 million reads for BA-411 and WI-6, respectively) of 150bp PE reads.

Reference-Guided Assembly

For each strain, we conducted reference-based assembly using the *D. pulex* PA42 genome assembly. *D. pulex* was determined to be a suitable, high-quality reference, as it is closely related to *D. pulicaria* and, interestingly, the two commonly hybridize in ecological communities (Marková *et al.* 2013; Kake-Guena *et al.* 2015). Furthermore, the PA42 genome was produced from starved *Daphnia* treated with antibiotics to reduce diet and endosymbiotic contaminants, and post-assembly, scaffolds were filtered for bacterial contamination (Ye *et al.* 2017). Our assembly pipeline, described below, was run on Auburn University's High-Performance Cluster, Hopper for 2 days using 20 cores and 100GB of memory.

For each strain, we used the following pipeline. Quality assessment of raw data files was performed using *FASTQC* v0.11.5 (Andrews, Simon 2010). The assessment reported no adapter contamination and no regions where sequence quality dropped below Q-score of 25, therefore trimming was not applied to reduce unnecessary loss of data. *D. pulicaria* reads were mapped to PA42 using Burrows-Wheeler Aligner (*BWA)* v0.7.15 (Li and Durbin 2009). Genome Analysis Tool Kit (*GATK)* v3.6 was used for local realignment, insertion/deletion (INDEL) and single nucleotide polymorphism

identification, and separation and filtration of identified variants using GATK recommended hard-filtering parameters (McKenna *et al.* 2010; Auwera *et al.* 2013). SNPs were inserted into the original reference, creating a consensus sequence, using

BCFTools 'consensus' (Li 2011). *BEDTools 'genomcov'* was used to create a BED file of regions lacking reference read coverage and *'maskfasta'* was used to mask the zero coverage and INDEL regions in the consensus sequence with "N's" (Quinlan and Hall 2010). This produced a reference-guided, draft genome assembly for each strain.

Assembly Metrics and Assessments

Although we starved the *Daphnia* before sequencing, it is likely there was still remnant algal cells and bacterial contaminates in our sequencing data. To identify these and any other contaminates, *BlobTools* (v 1.0) workflow A was used to quantify and visualize represented taxa, therefore identifying contamination from other phyla in the raw reads and the draft assemblies (Laetsch and Blaxter 2017). *FastQ Screen* was also used as a screening method for contaminants by mapping a subset of read libraries to a search library with *bowtie2 (Langmead and Salzberg 2012; Wingett and Andrews 2018)*. The search library genomes included with the program were used and genomes for PA42, the *D. pulex* mitochondria (PRJNA11866), and the green algae *Monoraphidium neglectum* (PRJNA293989), a close relative of the food source for *Daphnia*, were added (Crease 1999; Bogen *et al.* 2013). Assembly completeness was estimated with

Benchmarking Universal Single-Copy Orthologs (*BUSCO*) v4.0.6 analysis using both the eukaryote_odb10 and arthropoda_odb10 databases (Simão *et al.* 2015; Waterhouse *et al.* 2018). *BEDTools 'coverage'* was also used to determine depth of coverage at genes annotated in PA42. *Sourmash* v4 uses a MinHash derived algorithm to estimate similarity of genomic sequences and was used here to make pairwise comparisons between draft and reference assemblies (Brown and Irber 2016). *Sourmash* was used to create DNA sketches, or hash sketches, from both the assemblies and the merged raw reads. These reduced sequence data representations can be rapidly compared for overlapping k-mer sized read content (overlapping k-mer space) using a Jaccard similarity coefficient, however it does not give information about genomic contiguity or structure. Based on recommendations in *sourmash* documentation, signatures were computed for k-mer sizes of 21,31, and 51 bp, to minimize false positives and maximize matches. The distance measure output from this method is highly correlated with the frequently used genetic distance measurement, average nucleotide identity (ANI) (Ondov *et al.* 2016).

Data Availability

Supplemental files can be found on GSA figshare (https://figshare.com/s/28a738d36fc93b619109). File S1 is a figure from *FastQ Screen* analysis. File S2 is a tarball containing the reference-guided assemblies for BA_411 and WI_6. Assembly files with "clean" appended to the name have been filtered for scaffolds without reference coverage. File S3 is a tarball containing *Blobtools* output. File S4 is a

tarball containing BUSCO outputs. All sequence data are available under the NCBI BioProject Accession PRJNA702463. Code used to perform the data analyses for this work can be found on GitHub (https://doi.org/10.5281/zenodo.4635402).

RESULTS & DISCUSSION

Assemblies

PA42 is a quality *D. pulex* genome consisting of approximately 156 megabase pairs (8.6% gaps) organized into 1822 scaffolds. The BA-411 sequencing library produced 54.8 million reads with 24.9% duplication, and the WI-6 library produced 56.1 million reads with 21.7% duplication. Using BWA, BA-411 and WI-6 reads were mapped to PA42 resulting in approximately 86% and 75% successfully mapped reads, respectively. Of the 1822 scaffolds making up the PA42 assembly, 21 (1.15%) and 12 (0.66%) reference scaffolds had no sequence coverage for BA-411 and WI-6, with average coverage depths of ~33X and ~29X, respectively for the rest of the assembly. Assembly metrics are compared in Table 1. Although *D. pulex* and *D. pulicaria* are closely related, the assemblies presented herein are reference-guided and regions of the genomes that are not truly syntenic between the species will be incorrect in these BA-411 and WI-6 draft assemblies.

Approximately 97% of genes annotated in PA42 had coverage from mapped BA-411 and WI-6 reads. The percentage of PA42 genes with coverage is represented by the

"Total" category in Figure 1.1. The percentage of PA42 genes with 5, 10, 15, and 20X average depth are also found in this figure. Interestingly, BA and WI-6 have a similar number of genes with coverage, but WI-6 has consistently fewer genes covered at average depths of 10x or higher.

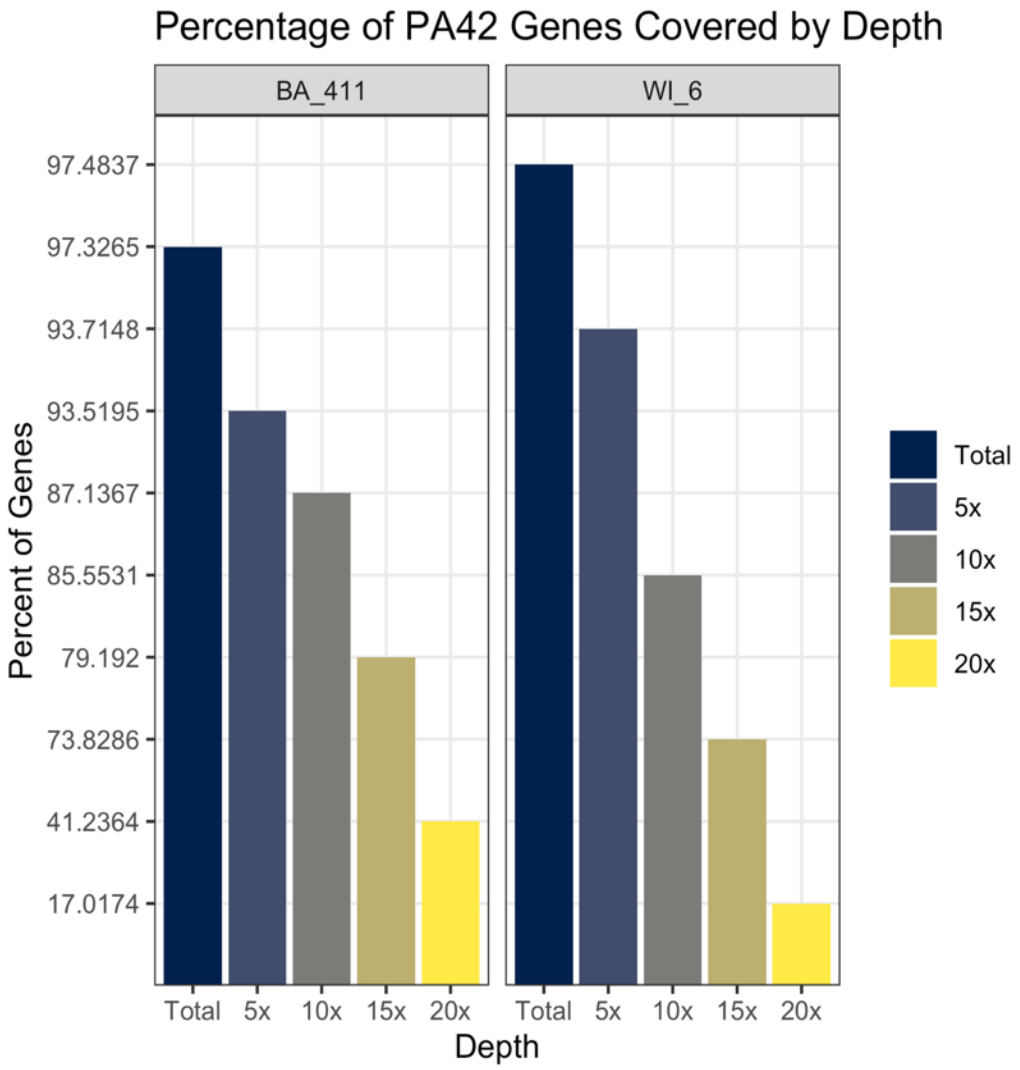

Figure 1.1. Percent of PA42 Genes for Different Average Depths of Coverage. The left and right panels represent BA-411 and WI-6, respectively. "Total" is the number of genes with 1x average depth. The percentage of genes covered at average depths of 5, 10,15, and 20x are included.

Assessments

To further assess the completeness of BA-411 and WI-6 assemblies, a *BUSCO* analysis was run using both the arthropod and eukaryote databases. Over 95% of the universal single-copy orthologs searched from both databases were found to be complete in both draft assemblies, with a minor difference in fragmented orthologs (0.1%; Figure 1.2A). The Venn diagram of missing BUSCOs in Figure 1.2B indicate that there are seven missing across all assemblies, corresponding to what is missing in the reference, and six BUSCOs that are missing in *D. pulicaria* only, with three species-specific orthologs missing in both strains and three strain-specific orthologs missing from each strain. These data indicate high contiguity in many genic regions for these draft assemblies.

Assemblies were assessed for contamination with *BlobTools*. We had an expectation of bacterial and algal contamination in the read data considering the microenvironment and diet of *Daphnia,* but because we used a reference sequence where great measures were taken to remove contaminants, we expected that a vast majority of contaminants would be filtered out during mapping. Based on the blob plots (Figure 1.3), both drafts genome assemblies had low levels of contaminant sequences, with 0.22% of BA-411 and 0.13% of WI-6 mapped reads hitting to phyla outside of Arthropoda. Supplementary data includes *BlobTools* output to further explore or remove contaminant regions.

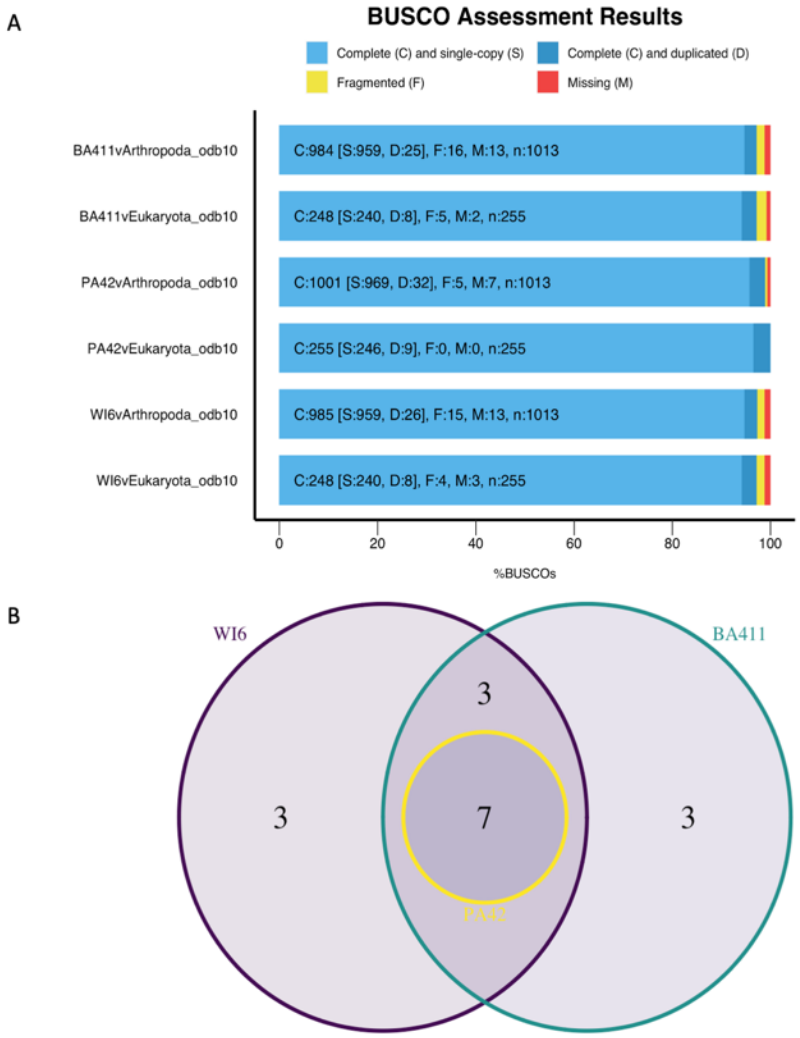

Figure 1.2. BUSCO Analysis for *D. pulicaria* Genome Assemblies Indicate High Levels of Gene Content in Draft Assemblies. A. BUSCO analysis for draft assemblies, BA-411 and WI-6, and the reference genome, PA42, against the eukaryote and arthropod databases. Colors indicate status of ortholog in the assembly. B. Venn diagram of missing arthropod BUSCOs for three *Daphnia* assemblies. Seven of the 13 missing BUSCOs in *D. pulicaria* assemblies were not present in the PA42 reference genome used for assembly.

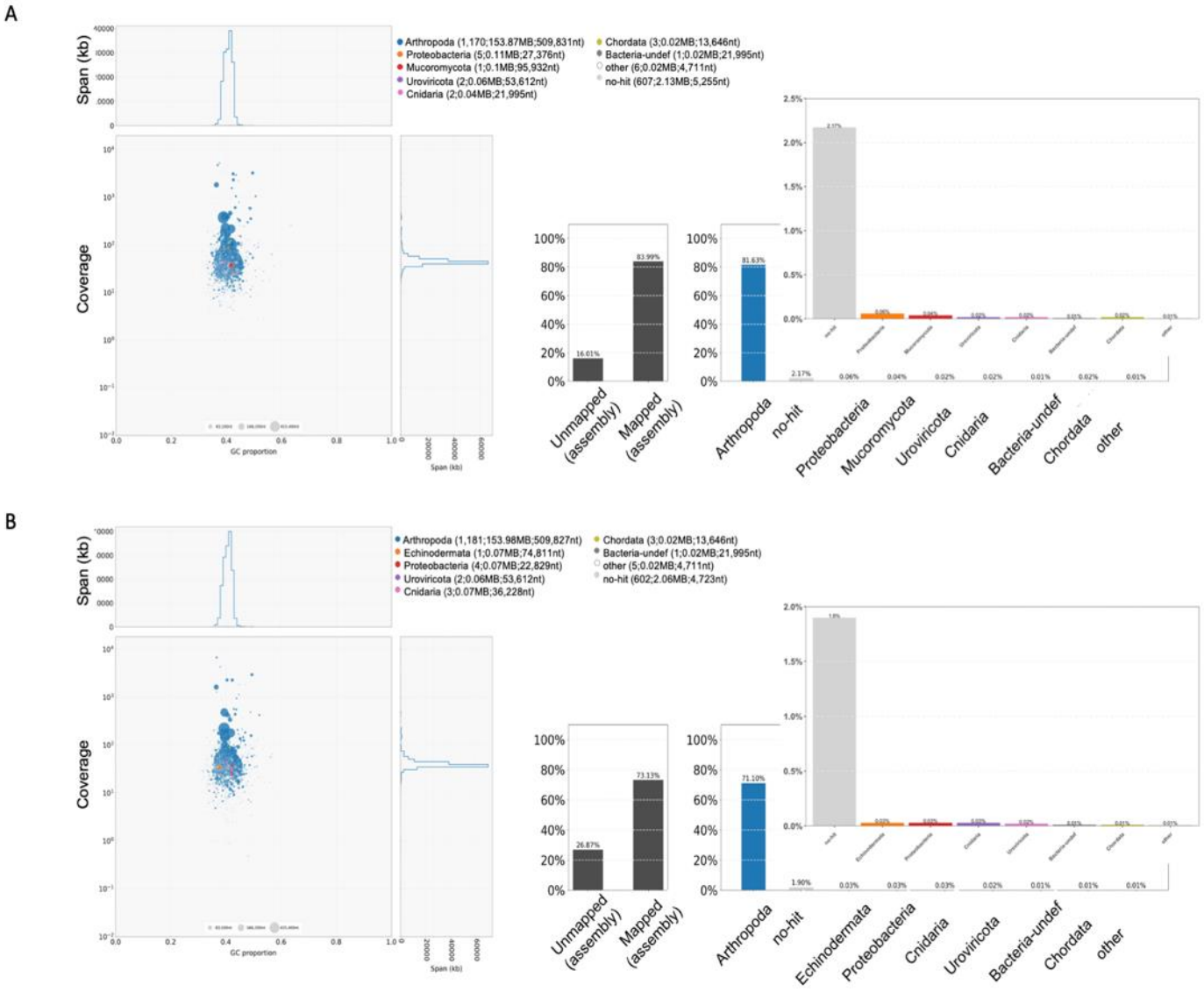

Figure 1.3: BlobPlots Indicate Low Levels of Contaminant Phyla in BA411 and WI6 Draft Reference-Guided Assemblies. Coverage by GC content scatterplots (BlobPlot) accompanied by read coverage plots for A. BA-411 and B. WI-6 draft assemblies. BlobPlots: The circles are the sequences, with sequence length proportional to circle diameter. The legend indicates each phyla represented with count, total span and N50 for each taxonomic rank in parentheses. Only a small number of sequences used in BLASTx analysis against the NCBI non-redundant protein database hit to phyla (19) other than the target Arthropoda. BarPlots: The grey bars represent the proportion of unmapped and mapped reads from libraries. Color bars represent the mapped proportion by taxonomic rank (phyla); an inset is included for viewing taxa present at low proportions.

FastQ Screen results (Supplementary Figure 1.1) corroborate the *BlobTools* analysis with a majority of the read subset mapping to the PA42 library. A portion of the read subsets (19-30%) mapped at low, non-specific levels or did not map at all to the other species and sequences included in the search library. This suggest that the appropriate search genome was not included, and it is likely that these reads may be unique to *D. pulicaria* or that a completely unexpected contaminate is present.

To gain a preliminary perspective on genetic distance between the two draft assemblies and the PA42 reference, we used *sourmash*. Distance estimates range from zero, being completely divergent to one, being completely identical. The assemblies from the *D. pulicaria* strains BA-411 and WI-6 had a computed distance of 0.90. This result is intuitive, as these are two strains of the same *Daphnia* species. BA-411 and WI-6 had very similar distance estimates for PA42 comparisons, with estimates of 0.73 for BA-411 and 0.74 for WI-6 (Supplementary Figure 1.2). These results corroborate the slight increase in gene content and PA42 scaffold coverage obtained from *BUSCO* analysis and mapping statistics for WI-6.

CONCLUSION

Daphnia species have long been studied in the context of ecology, evolution, and applied research. Here we present draft genome assemblies for two strains of *Daphnia* that vary in their tolerance to cyanobacteria. Algal blooms, characterized by an overgrowth (bloom) of cyanobacteria, are detrimental to the health of aquatic and terrestrial members of freshwater ecosystems. Population expansion of cyanobacteria is

caused by eutrophication, or the overloading of nutrients (e.g., phosphorus and nitrogen) in lakes, ponds and rivers, and is accelerated by increasing temperatures (Carpenter 2005; Schmale *et al.* 2019). From an economic perspective, algal blooms decrease water quality due to decreases in available oxygen and increases in toxic metabolites produced by cyanobacteria that result in product losses in fisheries, and toxification of water sources used by wild, domestic, and human populations for consumption and aquatic recreation (Anderson *et al.* 2002; Schmale *et al.* 2019). *Microcystis aeruginosa* is a highly toxic, cosmopolitan species of cyanobacteria that may produce metabolites called microcystins, compounds demonstrated to have significant hepatotoxic and tumorigenic effects (Paerl *et al.* 2001). Keeping levels of these damaging algal blooms in check is a particularly important and active branch of ecological research. Methods proposed for managing cyanobacteria include reducing the introduction of extraneous nutrients often from human runoff, the introduction of herbicide, and biomanipulation, or the manipulation of trophic levels to control cyanobacteria populations. Introducing toxin-tolerant *Daphnia pulicaria* has been shown to repeatedly lead to significant reductions of total algal biomass, including cyanobacteria, in limnocorral experiments (Wilson and Chislock 2013; Chislock *et al.* 2019a, 2019b).

In addition to understanding *D. pulicaria's* top-down regulation of algal biomass through mesocosm experiments, we are building resources to understand the genetic mechanisms and associations of toxic prey resistance observed in this species with these draft assemblies. Genomic resources are key components to deepening our understanding of the contributions of genetic background on strain-specific responses to toxic algal

blooms and other environmental stressors. These resources can be used for understanding the transcriptomic responses to toxins (Asselman *et al.* 2012; Orsini *et al.* 2016; Giraudo *et al.* 2017), identifying sequence variants under positive selection across the genome (Bourgeois *et al.* 2017; Schwarzenberger *et al.* 2020), and comparative analysis across other *Daphnia* species (Ravindran *et al.* 2019). In this way, these genomic resources provide a promising avenue for future research as the effects of urbanization and global climate change continue to exacerbate the severity of these toxic algal blooms over time (Carpenter 2005; Schmale *et al.* 2019).

These are reference-based *D. pulicaria* draft genome assemblies. In this study, 14–25% of the reads did not map to *D. pulex* PA42 genome assembly in our mapping. Similar to the read mapping percentages reported here (75–86%), Lack et al. (2018) produced sequencing libraries for pooled and individual D. pulicaria adults and ephippia and reported an average of ~72% mapping success to the TCO reference genome across 11 libraries (Lack *et al.* 2018). This indicates room for improvement in our assemblies. The data presented here are short-read sequences (150 bp paired-end). Future analyses should include long-read sequence data appropriate for de novo assembly that could recover the unmapped regions, improve scaffolds presented here, identify novel D. pulicaria scaffolds and chromosomal rearrangements to resolve conflicts in genetic structure between D. pulex and D. pulicaria genomes. Even with the aforementioned caveats of the two D. pulicaria genome assemblies we present, these assemblies contain very low levels of contamination, and high levels of genic content with more than 95% of complete universal arthropod and eukaryote orthologs found in these assemblies. This

work contributes quality reference-guided assemblies for two strains, one tolerant and one sensitive, of D. pulicaria that can be useful resources in linking candidate genes involved in ecologically relevant trait divergence, such as the evolution of dietary tolerance to toxic cyanobacteria, that impact freshwater communities and ecosystems.

ACKNOWLEDGEMENTS

We acknowledge the Auburn University Hopper Cluster for support of this work.

FUNDING

Financial support for this work was provided by an NSF GRF (1414475) to AC, by an NSF-REU fellowship in Computational Biology (1560115) to BH, by an NSF-REU grant in Aquatic Ecology (1658694) to AW, and start-up funds from Auburn University to TSS.

CONFLICT OF INTEREST

We declare no conflict of interest.

REFERENCES

Anderson, D. M., P. M. Glibert, and J. M. Burkholder, 2002 Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. Estuaries 25: 704–726. Andrews, Simon, 2010 FASTQC. A quality control tool for high throughput sequence data.

Asselman, J., D. I. M. De Coninck, S. Glaholt, J. K. Colbourne, C. R. Janssen *et al.*, 2012 Identification of pathways, gene networks, and paralogous gene families in *Daphnia pulex* responding to exposure to the toxic cyanobacterium *Microcystis aeruginosa*. Environ. Sci. Technol. 46: 8448–8457.

Asselman, J., J. D. Hochmuth, and K. A. C. De Schamphelaere, 2014 A comparison of the sensitivities of *Daphnia magna* and *Daphnia pulex* to six different cyanobacteria. Harmful Algae 39: 1–7.

Asselman, J., M. E. Pfrender, J. A. Lopez, J. R. Shaw, and K. A. C. De Schamphelaere, 2018 Gene coexpression networks drive and predict reproductive effects in *Daphnia* in response to environmental disturbances. Environ. Sci. Technol. 52: 317–326.

Auwera, G. A. V. der, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel *et al.*, 2013 From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics 43: 11.10.1-11.10.33. Becker, D., Y. Reydelet, J. A. Lopez, C. Jackson, J. K. Colbourne *et al.*, 2018 The transcriptomic and

proteomic responses of *Daphnia pulex* to changes in temperature and food supply comprise environment-specific and clone-specific elements. BMC Genomics 19: 376.

Bogen, C., A. Al-Dilaimi, A. Albersmeier, J. Wichmann, M. Grundmann *et al.*, 2013 Reconstruction of the lipid metabolism for the microalga *Monoraphidium neglectum* from its genome sequence reveals characteristics suitable for biofuel production. BMC Genomics 14: 926.

Bourgeois, Y., A. C. Roulin, K. Müller, and D. Ebert, 2017 Parasitism drives host genome evolution: Insights from the *Pasteuria ramosa–Daphnia magna* system. Evolution 71: 1106–1113.

Brown, C. T., and L. Irber, 2016 sourmash: a library for MinHash sketching of DNA. Journal of Open Source Software 1: 27.

Brunner, F. S., J. A. Deere, M. Egas, C. Eizaguirre, and J. A. M. Raeymaekers, 2019 The diversity of eco-evolutionary dynamics: Comparing the feedbacks between ecology and evolution across scales. Functional Ecology 33: 7–12.

Carpenter, S. R., 2005 Eutrophication of aquatic ecosystems: Bistability and soil phosphorus. PNAS 102: 10002–10005.

Chislock, M. F., R. B. Kaul, K. A. Durham, O. Sarnelle, and A. E. Wilson, 2019a Eutrophication mediates rapid clonal evolution in *Daphnia pulicaria*. Freshw Biol 64: 1275–1283.

Chislock, M. F., O. Sarnelle, L. M. Jernigan, V. R. Anderson, A. Abebe *et al.*, 2019b Consumer adaptation mediates top–down regulation across a productivity gradient. Oecologia 190: 195–205.

Colbourne, J. K., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker *et al.*, 2011 The ecoresponsive genome of *Daphnia pulex*. Science 331: 555–561.

Colbourne, J. K., B. Robison, K. Bogart, and M. Lynch, 2004 Five hundred and twentyeight microsatellite markers for ecological genomic investigations using *Daphnia*. Molecular Ecology Notes 4: 485–490.

Colbourne, J. K., V. R. Singan, and D. G. Gilbert, 2005 wFleaBase: the *Daphnia* genome database. BMC Bioinformatics 6: 45.

Crease, T. J., 1999 The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). Gene 233: 89–99.

Eads, B. D., J. Andrews, and J. K. Colbourne, 2008 Ecological genomics in *Daphnia*: stress responses and environmental sex determination. Heredity 100: 184–190. Ebert, D., 2005 *Introduction to Daphnia Biology*. National Center for Biotechnology Information (US).

Giraudo, M., M. Douville, G. Cottin, and M. Houde, 2017 Transcriptomic, cellular and life-history responses of *Daphnia magna* chronically exposed to benzotriazoles: Endocrine-disrupting potential and molting effects. PLOS ONE 12: e0171763.

Hairston, N. G., C. L. Holtmeier, W. Lampert, L. J. Weider, D. M. Post *et al.*, 2001 Natural selection for grazer resistance to toxic cyanobacteria: Evolution of phenotypic plasticity? Evolution 55: 2203–2214.

Jia, J., G. Hu, C. Feng, C. Dong, M. Han *et al.*, 2020 *Daphnia carinata* genome provides insights into reproductive switching: Preprints preprint.

Kake-Guena, S. A., K. Touisse, R. Vergilino, F. Dufresne, P. U. Blier *et al.*, 2015 Assessment of mitochondrial functions in *Daphnia pulex* clones using high-resolution respirometry. Journal of Experimental Zoology Part A: Ecological Genetics and Physiology 323: 292–300.

Lack, J. B., L. J. Weider, and P. D. Jeyasingh, 2018 Whole genome amplification and sequencing of a *Daphnia* resting egg. Molecular Ecology Resources 18: 118–127.

Laetsch, D. R., and M. L. Blaxter, 2017 BlobTools: Interrogation of genome assemblies. F1000Res 6: 1287.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Lee, B.-Y., B.-S. Choi, M.-S. Kim, J. C. Park, C.-B. Jeong *et al.*, 2019 The genome of the freshwater water flea *Daphnia magna*: A potential use for freshwater molecular ecotoxicology. Aquatic Toxicology 210: 69–84.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–

Wheeler transform. Bioinformatics 25: 1754–1760.

Marková, S., F. Dufresne, M. Manca, and P. Kotlík, 2013 Mitochondrial capture misleads about ecological speciation in the *Daphnia pulex* complex. PLOS ONE 8: e69497.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Miner, B. E., L. De Meester, M. E. Pfrender, W. Lampert, and N. G. Hairston, 2012

Linking genes to communities and ecosystems: *Daphnia* as an ecogenomic model.

Proceedings of the Royal Society B: Biological Sciences 279: 1873–1882.

Nelson, J. R., T. S. Schwartz, and J. M. Gohlke, 2018 Influence of maternal age on the effects of seleno-l-methionine in the model organism *Daphnia pulex* under standard and heat stress conditions. Reproductive Toxicology 75: 1–9.

Nickel, J., T. Schell, T. Holtzem, A. Thielsch, S. R. Dennis *et al.*, 2021 Hybridization dynamics and extensive introgression in the *Daphnia longispina* species complex: new insights from a high-quality *Daphnia galeata* reference genome. bioRxiv

2021.02.01.429177.

Nobel Media AB 2021 Ilya Mechnikov – Facts. NobelPrize.org.

Ondov, B. D., T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman *et al.*, 2016

Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17: 132.

Orsini, L., D. Gilbert, R. Podicheti, M. Jansen, J. B. Brown *et al.*, 2016 *Daphnia magna* transcriptome by RNA-Seq across 12 environmental stressors. Sci Data 3: 160030. Paerl, H. W., R. S. Fulton, P. H. Moisander, and J. Dyble, 2001 Harmful freshwater algal blooms, with an emphasis on cyanobacteria. The Scientific World JOURNAL 1: 76–113. Pelletier, F., D. Garant, and A. P. Hendry, 2009 Eco-evolutionary dynamics. Philos Trans R Soc Lond B Biol Sci 364: 1483–1489.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Ravindran, S. P., J. Lüneburg, L. Gottschlich, V. Tams, and M. Cordellier, 2019 *Daphnia* stressor database: Taking advantage of a decade of *Daphnia* '-omics' data for gene annotation. Sci Rep 9: 11135.

Sarnelle, O., S. Gustafsson, and L.-A. Hansson, 2010 Effects of cyanobacteria on fitness components of the herbivore *Daphnia*. Journal of Plankton Research 32: 471–477. Sarnelle, O., and A. E. Wilson, 2005 Local adaptation of *Daphnia pulicaria* to toxic cyanobacteria. Limnol. Oceanogr. 50: 1565–1570.

Schmale, D. G. I., A. P. Ault, W. Saad, D. T. Scott, and J. A. Westrick, 2019 Perspectives on harmful algal blooms (HABs) and the cyberbiosecurity of freshwater systems. Front. Bioeng. Biotechnol. 7:128

Schwarzenberger, A., M. Hasselmann, and E. V. Elert, 2020 Positive selection of digestive proteases in *Daphnia*: A mechanism for local adaptation to cyanobacterial protease inhibitors. Molecular Ecology 29: 912–919.

Shaw, J. R., J. K. Colbourne, J. C. Davey, S. P. Glaholt, T. H. Hampton *et al.*, 2007 Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics 8: 477.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35: 543–548.

Wilson, A. E., and M. F. Chislock, 2013 Ecological control of cyanobacterial blooms in freshwater ecosystems, pp. 213–221 in *Cyanobacteria: Ecology, Toxicology and Management*, edited by Ferrão-Filho, Aloysio Da S. Nova Science Publishers, Inc, New York, NY.

Wilson, A. E., and M. E. Hay, 2007 A direct test of cyanobacterial chemical defense: Variable effects of microcystin-treated food on two *Daphnia pulicaria* clones. Limnology and Oceanography 52: 1467–1479.

Wingett, S. W., and S. Andrews, 2018 FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res 7: 1338.

Ye, Z., S. Xu, K. Spitze, J. Asselman, X. Jiang *et al.*, 2017 A cew reference genome assembly for the microcrustacean *Daphnia pulex*. G3 (Bethesda) 7: 1405–1416.

Table 1.1. A table of statistics from reference-guided assemblies for *Daphnia pulicaria*. Mapping and base calling statistics for *D. pulicaria* sequencing libraries from this study.

LOCATION	LIBRARY	GDNA CONC. [ng/ml]	READS GENERATED	READS MAPPED	% READS MAPPED	MEAN DEPTH	CALLED SITES	N's	PA42 Mb COVERED	SRA ACCESSION
Bassett Lake	USD16091408	28	54849146	47627534	86.83	33.55	2923235	28692854	140	SRR14023941
Wintergreen Lake	USD16091409	21.6	56113536	42293065	75.37	29.46	3070431	27250307	141	SRR14023940

CHAPTER 2:

Comparison of 18 common RNA-seq pipelines for differential gene expression analysis: from mapping to functional pathway enrichment

Keywords**:** RNA-Seq, Differentially expressed genes; Functional genomics,

This work will contribute to three manuscripts in preparation:

Clark, AD, S Seungyeon, M Khan, B Sipley, A Pokhrel, R Telemeco, D Waites, C Ugochukwu, S Tansie, TS Schwartz. Comparison of 18 common RNA-seq pipelines for differential gene expression analysis: from mapping to functional pathway enrichment.

Schwartz, TS, AD Clark, P Pearson, J Roberts, J Dawson, DB Allison, J Gohlke. Genetic response of dietary restriction extending lifespan in a clone of *Daphnia pulex*.

Schwartz TS and AD Clark. Structure for a Functional Genomics CURE using RNAseq Analysis. Planned submission to CourseSource.

INTRODUCTION

Characterizing transcriptional profiles generated in response to a range of conditions, treatments, timepoints, and other biologically relevant states by sequencing cDNA reverse transcribed from mRNA (RNA-sequencing or RNA-seq) is now a standard molecular practice. Often the end goal for generating transcriptional profiles across these states is to test for differential gene expression among groups (e.g. drug treatments, environmental stressors, disease states, etc.) to understand the biological functions that underly the phenotypic differences between the groups. Concomitant with the steady decrease in costs of high throughput sequencing, we have increasingly generated observations of the nature of RNA-seq data across the Tree of Life and we have developed a variety of methods to handle the complex nature of these multi-dimensional data (Conesa *et al.* 2016; Koch *et al.* 2018). This creates a challenge in understanding which methods to use for analysis, the degree to which these choices matter for biological interpretation, and how to train students to effectively conduct these analyses.

The classic RNA-seq analysis for the purpose of testing for differential gene expression consists of the steps outlined in Figure 2.1. Briefly, raw sequencing reads are aligned (mapped) to targeted genomic loci, most often genome assemblies and/or transcripts, to obtain the number of reads sequenced (counts) from mRNA that were transcribed from those loci, followed by gene-wise tests for statistically differential gene expression between samples structured by groups of comparative interest and normalized for technical and inherent biological variation. Frequently in the absence of *a priori* expectations of which genes will be differentially expressed, the next step is testing for

the enrichment of functional pathways amongst the top differentially expressed genes (Kukurba and Montgomery 2015; Stark *et al.* 2019). Although RNA-seq methods have become a standard practice for quantifying gene expression, analysis of these data is considerably variable, with new programs continually being developed to improve read mapping and counting, statistical models, and model assumptions employed during differential gene expression (DGE) analysis. Mixing and matching the programs across these steps can produce hundreds of potential pipelines. This leaves researchers a plethora of decisions to make about the appropriate programs for their dataset at each step of the analysis and, importantly, the degree to which the choice of the programs at each of steps matters in the context of the biological interpretation of the experiment.

Systematic comparisons of programs at individual steps have been carried out by several other researchers and the authors creating them, with the purpose of benchmarking their new algorithms with existing ones as a demonstration of similar or better performance (Engström *et al.* 2013; Soneson and Delorenzi 2013; Love *et al.* 2014; Seyednasrollah *et al.* 2015; Baruzzo *et al.* 2017; Abrams *et al.* 2019). Srivastava et al (2020) recently compared alignment-based and alignment-free methods using mouse data to understand variation in tools aligning to the genome, the transcriptome or pseudoaligning to the transcriptome. Their work also identified shortcomings with using simulated data in benchmarking RNA-seq pipeline tools. Simulated reads lack natural variation that exists between a transcriptome and the reads aligned to that transcriptome during quantification. They also lack the of complexities of real read libraries which have more variation in composition (e.g., alternative splicing, intronic or intergenic sequences). Previous studies with simulated data were not able to fully capture some of

the impactful effects of spurious mappings produced by different alignment methods that propagated to downstream quantification and detection of differentially expressed genes (DEGs) (Srivastava *et al.* 2020). RNA-seq analysis pipeline comparisons at the step of alignment programs have also been made using a highly polymorphic species. Schaarschmidt et al. (2021) compared two accessions of the plant model, *Arabidopsis thalania,* in ambient (20^{o}C) and cold (4^{o}C) conditions and found small differences in mappability across aligners, high similarity in raw read counts across all aligners & pseudoaligners (Schaarschmidt *et al.* 2020). There was more variability (92 – 98%) in similarity between programs used in the detection of differentially expressed genes (DEGs), specifically when comparing DEG results between the program DESeq2 (Love *et al.* 2014) and CLC (Qiagen). More complex, multi-step comparisons of full pipelines, from aligners to DGE programs, have been done with cancer cell line responses to two different therapeutic drugs in Corchete et al. (2020). Their work tested all possible combinations of 3 trimmers, 5 aligners, 6 counters, 3 pseudoaligners, 8 normalization methods, and 17 DGE programs for the best ranked pipelines. This ranking was based on a combination of precision and accuracy relative to gold-standard qRT-PCR expression data (Ct) for a subset of genes ubiquitously expressed across healthy, control samples. While their work found variation in precision and accuracy across methods at each step of analysis, the largest, statistically significant differences were found in counting and normalization methods (Corchete *et al.* 2020).

These invaluable studies have aided in our understanding of the effects of different algorithms and assumptions implemented in programs used at different steps of RNA-seq analysis, with some estimations of accuracy and precision when qRT-PCR data

could be generated. Still, there have been limited comparisons that span alignment to functional pathway enrichment for biological interpretation to determine if, at the end of the analyses, there would be similar (or very different) biological conclusions based on which turn in the pipeline maze an investigator has made. In the interest of bringing these underlying technical conclusions about variation in pipeline tools back to the ultimate perspective of biological interpretation, we quantify similarities in functional enrichment, mapping, counting, and DGE results from combinations of 2 aligners, 2 counters, 2 pseudoaligners, 3 DGE normalization and detection programs, and one final program for functional enrichment, for a total of 18 bioinformatics pipelines.

In contrasting these 18 bioinformatics pipelines for similarities in biological interpretation, we are also addressing a fundamentally important biological question on how caloric restriction extends lifespan using a non-model organism relevant in ecotoxicology and evolution, *Daphnia pulex*. Caloric restriction has been shown to increase lifespan across the animal kingdom, from yeast to rodents (Osborne *et al.* 1917; Anderson and Weindruch 2010). We propose *Daphnia* as a complementary model for understanding variation in lifespan due to caloric restriction from both an evolutionary perspective and in its ability to translate results from laboratory and natural populations. *Daphnia* are short lived (median 1 mo.) and have a similar mortality curve as mammals (Jones *et al.* 2014). Their genome is stable due to their clonality, and their relatively small genomes are more similar to humans than *Drosophila* or *Xenopus* (Colbourne *et al.* 2011b).

Thus, the goals of this chapter are 3-fold. We aim to answer the following questions: (1) How do different pipelines vary in the final functional/biological

interpretations made? (2) What biological pathways are affected by caloric restriction in *Daphnia pulex*? Lastly, we aim to develop reproducible, open-source code available via GitHub along with tutorials for instructors that will be used for teaching RNA-seq analysis.

MATERIALS & METHODS

The phenotypic experiment and resulting RNA-sequencing was conducted by Tonia Schwartz at University of Alabama at Birmingham in 2014 and will briefly be described here to provide the appropriate background information for the RNA-seq analysis herein. **The RNA-seq data analysis and biological interpretation was conducted by Amanda Clark at Auburn University.**

Caloric Restriction Experiment

The experiment that generated these RNA-seq data was performed in 2014 and used a strain of *Daphnia pulex* maintained in Dr. Julia Gohlke's laboratory at University of Alabama at Birmingham since 2011. *Daphnia* were maintained in COMBO media and fed RGcomplete [Reed Mariculture], which is a blend of four microalgae (1.5 – 15 µc). The diet treatments and media used in this experiment were defined in Schwartz et al. (2016): C-treatment represents the caloric restriction treatment with 98 µl of RGcomplete per liter of COMBO media, and the E-treatment represents the *ad libitum* treatment with 300 µl of algae per liter of COMBO media. Populations were maintained at the high food concentration (E treatment) for two generations prior to the third

generation of neonates being randomly assigned into either C or E populations. Each treatment had eight populations (1 liter beaker) of 20 individuals with 50 ml of media/individual.

To provide insight into the gene regulatory mechanisms associated with the extension of lifespan in response to caloric restriction in *Daphnia*, animals were collected from C and E populations at 23 days of age for transcriptomic analysis using RNA-seq. For each diet treatment, 5 populations were randomly sampled for 3 individuals. The five samples from each treatment were dissected to remove offspring from their brooding pouches and immersed in RNAlater [Qiagen] for two days at 4^{o}C before RNA isolation (n=10 samples each with 3 individuals). *Daphnia* were removed from the RNAlater, quickly rinsed in water, and snap frozen in liquid nitrogen in a 1.5ml tube for homogenization by pestle. Total RNA was isolated using the RNeasy Mini Kit [Qiagen] with DNA digestion on the membrane.

Library Preparation and RNA-seq Sequencing

RNA samples were sent to the Heflin Genomic Center at University of Alabama for sequencing using the Illumina HiSeq2500 (Illumina, San Diego, CA) and the Agilent SureSelect Stranded library preparation kit (Agilent Technologies, Santa Clara, CA). Quality and quantity of RNA were determined on the Bioanalyzer. 100ng of total RNA was subjected to two rounds of poly A+ selection using oligo dT magnetic beads. Following purification, the mRNA was randomly fragmented, and first strand cDNA synthesis was done in the presence of random hexamers and 2.4ng/µL (final concentration) of Actinomycin D using standard techniques. First strand cDNA was

purified by magnetic bead (Omega Bio-Tek, Norcross, GA) prior to second strand synthesis. After second strand synthesis was complete the cDNA was adenylated and used in a ligation reaction to add primary adaptors. Final libraries were purified by magnetic beads, quantitated using the KAPA SYBR FAST qPCR kit (KapaBiosystems, Woburn, MA) on the Roche LightCycler 480 (Roche, Indianapolis, IN) and assessed for quality on the High Sensitivity DNA chip for the Agilent BioAnalyzer (Agilent Technologies, Santa Clara, CA). Sequencing libraries were mixed to equal molar amounts and run on the HiSeq2500 using a Rapid Run flow cell with paired end 100bp sequencing reads. Following completion of the run the .bcl files were converted to FASTQ file format using BCL2FASTQ 1.8.4 from Illumina. Libraries were sequenced (100 bp paired-end) on a single rapid run flow cell on the Illumina HiSeq 2500, with the 10 libraries split among two lanes (5 libraries multiplexed per lane). Data were submitted to NCBI SRA database under Bioproject PRJNA437447.

RNA-seq Analysis Tools by Step

Here we briefly describe the tools being compared at each step in a typical RNA-seq analysis pipeline. The tools are illustrated in Figure 2.1 and descriptions, versions, and associated parameters used in the bioinformatic pipelines are listed in Table 2.1.

Quality Assessment & Trimming

For these analyses, raw reads were downloaded with SRA Toolkit (SRA Toolkit Development Team) from NCBI (bioproject number: PRJNA437447). Quality of the reads were assessed using FastQC (Babraham Bioinformatics) (Andrews, Simon 2010).

Reads were trimmed and filtered using paired-end parameters in Trimmomatic (Bolger *et al.* 2014). The first ten base pairs and reads with a quality cutoff of below 30 were removed. Reads below the minimum length of 36 base pairs were removed and quality was assessed again using FastQC. Around 10% of reads were removed from the raw data and high quality was reported across the samples.

Alignment Programs

Two genome-based alignment methods, Hisat2 (Kim *et al.* 2019) and STAR (Dobin *et al.* 2013) were used to map trimmed reads to the PA42 *Daphnia pulex* genome as a reference genome (Ye *et al.* 2017). Annotation information (GTF) was provided to both aligners during index generation to take advantage of spliced aligners for mapping to a genome. Parameter specifics and tool descriptions are outlined in Table 2.1.

Pseudoalignment Programs

Recently, tools have been developed for rapid quantification of transcripts that do not generate full read alignments and combines mapping, counting, and normalization steps in a single program. Here we use two well-known quasi-mapping RNA-seq quantification programs, Salmon (Patro *et al.* 2017) and Kallisto (Bray *et al.* 2016). Parameter specifics and tool descriptions are outlined in Table 2.1.

Counting & Gene-Level Count Estimation Programs

For pipelines starting with traditional aligners, two programs, StringTie (Pertea *et al.* 2015) and HTSeq (Anders *et al.* 2015), were used to quantify reads overlapping targeted genomic loci, specifically genes. HTSeq-count automatically generates gene-

level abundance estimates by counting reads assigned to a feature. Pseudoaligners and pipelines using StringTie estimate individual transcript abundance, therefore the R package tximport (Soneson *et al.* 2016) was used to estimate gene-level abundance from transcript abundance. Parameter specifics and tool descriptions are outlined in Table 2.1.

All remaining analyses were carried out in R, where packages were obtained from CRAN unless specified as Bioconductor packages.

Pre-filtering Low and No-Expression Genes

In the interest of maintaining our systematic approach to pipeline comparison, we pre-filtered genes based on having low gene-level count estimates from all pipelines and disabled any downstream filtering within individual DGE analysis programs. Three filtering approaches were used to remove no and low count genes with the intent of increasing power to detect DEGs by reducing the number of statistical tests (Bourgon *et al.* 2010). First, counts from each pipeline were filtered individually, removing any gene with zero counts in 6 or more of the 10 samples and any gene with less than 21 counts across all samples. These datasets were labeled "pipeline_filtered." The other two filtering methods standardize the number of genes going into downstream analyses while using the previous filtering logic. Specifically, genes that would be filtered from any of the pipelines were removed from all pipelines (compilation), and generated datasets labeled "hard_filtered" or genes that would be filtered from all pipelines (intersection) were removed from all pipelines and generated datasets labeled "soft_filtered." We report results across filtering methods for select analyses, and others we prioritize the results from a single method for more relevant interpretation and application.

Differential Gene Expression (DGE) Programs

We use three different programs for DGE analysis with different methods of normalization and modeling approaches. We outline parameters and commands used during analyses for each program below but see Table 2.1 for parameter specifics and tool descriptions not discussed herein.

DESeq2

All default parameters were used apart from the parameters to filter lowly expressed genes, which were not used. minReplicatesForReplace, this parameter is used to denote the minimum number of replicates required to replace outliers in a sample which was set to "Inf" to never replace outliers. independentFiltering, DESeq2 package performs independent filtering of count data by default using mean of normalized counts. Since our gene count data was prefiltered we disabled this option (independentFiltering, = FALSE). "cooksCutoff" is used to set threshold to define outlier to be replaced. DESeq2 automatically flags the genes which have Cook's distance above a cutoff for samples that have 2 or more replicates. Since our data was prefiltered we disabled this option (cooksCutoff = FALSE).

EdgeR

We created a DGEList object to store gene-level counts and hold associated metadata using EdgeR (Robinson *et al.* 2010), grouping our gene count data based on treatment. We then normalized within/between samples using the (default) trimmed mean of M-values (TMM) method within the function "calcNormFactors." Next, we estimated

tagwise dispersion using "estimateTagwis eDisp" and performed an exact test to compare the ad lib and caloric restricted groups using default settings. Finally, we extracted significantly differentially expressed genes using the "topTags" function, keeping only the genes with FDR below 0.05. We did not perform any additional filtering using EdgeR.

Limma-Voom

We created a DGEList object to store gene-level counts and hold associated metadata using EdgeR, grouping our gene count data based on treatment. We then calculated normalization factors using the function calcNormFactors(DGEList, method="TMM") as with the EdgeR analysis. We specified the model using treatment as the predictor variable. Using the model residuals, voom (Law *et al.* 2014) estimates variance weights on a per observation basis (gene and sample-wise) using transformed counts with the normalization factors calculated in EdgeR. These variance weights are used with transformed counts in the standard linear models in Limma (Ritchie *et al.* 2015).

Pathway Analysis Program

Gene Set Enrichment Analysis (GSEA) was implemented in the fgsea package using pre-ranked gene generated for all pipelines (Korotkevich *et al.* 2021). Output tables from all DE programs included (1) p-values from t-tests of differential expression between treatments, (2) effects sizes and direction reported in log fold-change, and (3) adjusted p-values estimated using Benjamini-Hochberg false-discovery rate correction for multiple hypothesis testing across genes for each gene ID. Gene ranks were calculated as

the signed $\log_{10}$ p-value (sign of the effect size * - $\log_{10}$ of the p-value) before use in fgsea, where sign indicated upregulation (+) or downregulation (-) in the caloric restriction group relative to the *ad lib* fed group (Plaisier *et al.* 2010; Reimand *et al.* 2019). The hallmark gene sets from the Molecular Signatures Database (MSigDB) were used in these analyses (Liberzon *et al.* 2015). An FDR for significant enrichment of a gene set was set to 0.25 to identify pathways that would be of interest from a "discovery" approach.

Pipeline Contrasts and Statistical Tests

We compare pipelines in multiple ways to understand the relative contributions of each program to the variance between analysis results. We compare the amount of uniquely mapped and unmapped reads between aligners for all samples to estimate the mappability of each read library. We estimate the similarity of raw count matrices (capturing variation due to aligner/counter combinations) using Spearman correlations in scatter matrices per sample using the GGally package. R_v coefficients, which are Pearson's correlations generalized for matrices (Josse and Holmes 2016), were calculated for transformed, raw and soft_filtered count matrices in pairwise pipeline combinations using FactoMineR and visualized with a heatmap generated with the Bioconductor package ComplexHeatmap.

We assessed the relative contribution to variance in the number of biologically and statistically relevant DEGs for different steps of RNA-seq analysis using two linear models. We specify the response variable as the number of DEGs that had an FDR less than 0.05 (statistical relevance) and had $\log_2$ fold change greater than or equal to an

absolute value of 2 (biological relevance). Our models tested whether, quantification (aligner and counter) and/or DGE program predicted our response variable. Described here is our final model, however we validated the superior fit of the model over generalized linear models (Poisson and negative binomial distributions) using Akaike's Information Criterion (AIC) comparison. For each model, variance partitioning was performed using the Anova function in the car package. Estimated marginal means and contrasts for *post hoc* analyses (tukey's adjustment) were performed using the emmeans package, but the marginal effects were visualized using the ggeffects package. In our first model, we explore the predictors of DGE program and the combination of aligner & counter programs as a single, second variable. This allows for the comparison of pseudoaligners, which perform both steps, with the other quantification combinations of aligners and counters. Our second model uses aligners, counters, and DGE programs as predictor variables, and excludes data from pseudoaligners so that we can estimate the effects of aligner and counter, separately. We estimate the similarity of biologically and statistically significant DEG lists across pipelines visualizing gene set intersections with upset plots using UpsetR.

We planned to assess these same effects at the level of functional enrichment analyses with the number of statistically significant (FDR less than 25%) enriched gene sets as the response variable. Upon completing GSEA, 78% of the pipelines (across filtering methods) did not return any pathway that fit our criteria. This was likely not driven by the gene expression data, but more by the completeness of our annotation, so we did not model these data.

RESULTS & DISCUSSION

RNA-seq workflow:

We compare results from 18 RNA-seq pipelines using data generated from *Daphnia* in caloric restricted or ad lib diet treatments. We use 6 combinations of different programs for alignment and counting. We also tested 3 programs for DGE combined with the normalization procedures available in their respective packages (Fig. 2.1).

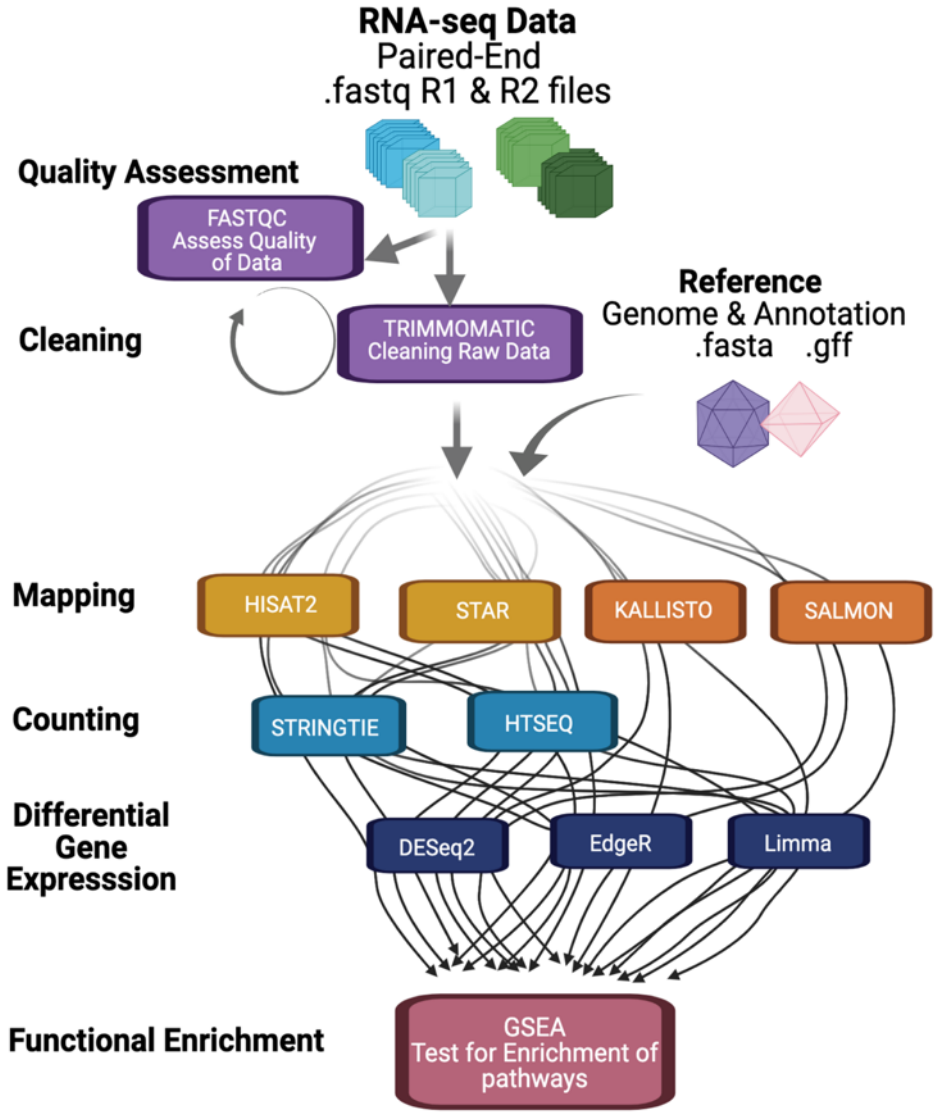

Figure 2.1: RNA-seq Pipelines for Differential Gene Expression Analysis Compared.

General RNA-seq analysis steps are on the left. At each step, the programs compared are in the center boxes. Small arrows are indicative of possible paths through these analysis programs. Single programs were used for Quality Assessment, Cleaning, and Functional Enrichment steps. Yellow and orange boxes at the alignment step differentiate traditional aligners and pseudoaligners, respectively. This figure was created with BioRender.com.

Aligners

Mapping percentages averaged across all samples for Hisat2 (58.4%) were lower than mapping percentages in STAR (67.97%) (Figure 2.2). These results are consistent with previous findings from other researchers that report higher mapping percentages in comparisons between these two alignment algorithms (Schaarschmidt *et al.* 2020; Musich *et al.* 2021). Interestingly, when averaging across biological replicates for treatment groups we see higher mapper percentages in control replicates (Hisat2 - 61.51%; STAR – 72.29%) than in diet restricted replicates (Hisat2 – 55.29%; STAR – 63.66%) across both alignment algorithms. This could indicate differential isoform usage in response to the caloric restriction treatment that is not annotated in our reference genome. This would not be a farfetched conclusion as the current annotation has one representative transcript for each gene feature.

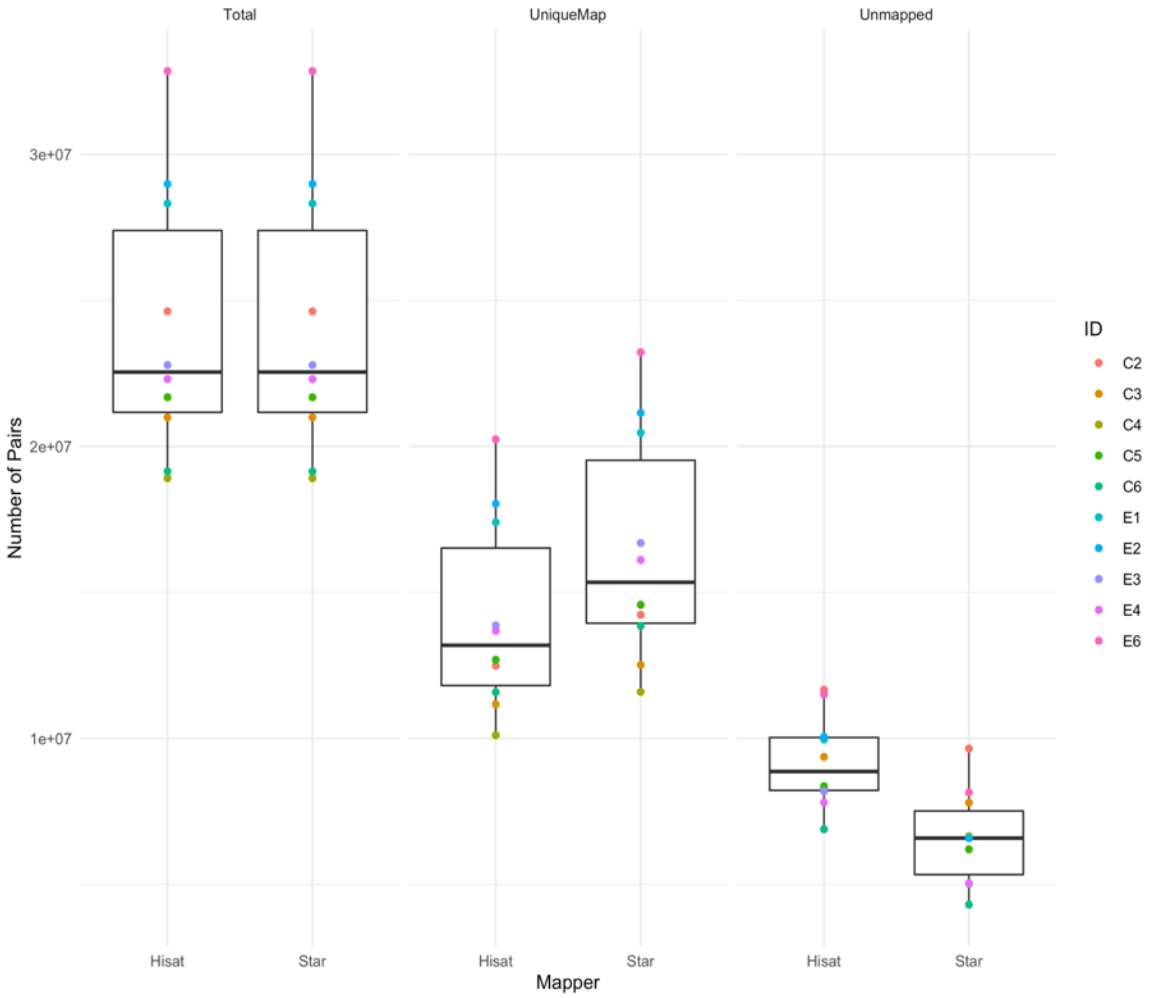

Figure 2.2: Mapping Statistics for Hisat2 and STAR aligners across samples.

Alignment data for the number of read pairs across alignment algorithms. This plot is faceted by the statistic (total number, uniquely mapped, and unmapped reads). "Total" is the total number of paired reads. "UniqueMap" are the number of read pairs that mapped to a single region of the genome. "Unmapped" are reads pairs that did not map to a region of the genome. "ID" are sample ideas from the caloric restriction experiment. IDs that begin with C are caloric restriction replications. IDs that begin with E are *ad lib* replicates.

Counting & Gene-Level Quantification

When contrasting aligners, we only compare Hisat2 and STAR because the two pseudoaligners do not produce mapping statistics. Although you can get "pseudobams" from Kallisto or use a different aligner prior to quantification in Salmon, we felt it would not provide clean or independent comparisons of these tools with true alignment algorithms. Considering this, we made comparisons at the level of quantification using 6 aligner/counter combinations. We find high correlations coefficients (0.946 – 0.982, all p-values < 0.001) across the pipelines at this level of comparison (Figure 2.3). Unsurprisingly, the lowest correlation coefficient of 0.946 was between a pseudoaligner (Kallisto) and traditional aligner/counter (Hisat2-StringTie). The second lowest correlation coefficient (0.948) was observed in the comparison between Hisat2-StringTie and STAR-HTSeq combinations where both the aligner and counter varied.

At the upper range of the correlation coefficients, 0.982 (both aligners + StringTie), 0.981 (both aligners + Hisat2), and 0.970 (both pseudoaligners), we see that correlation estimates are highest when the methods for counting are held constant.

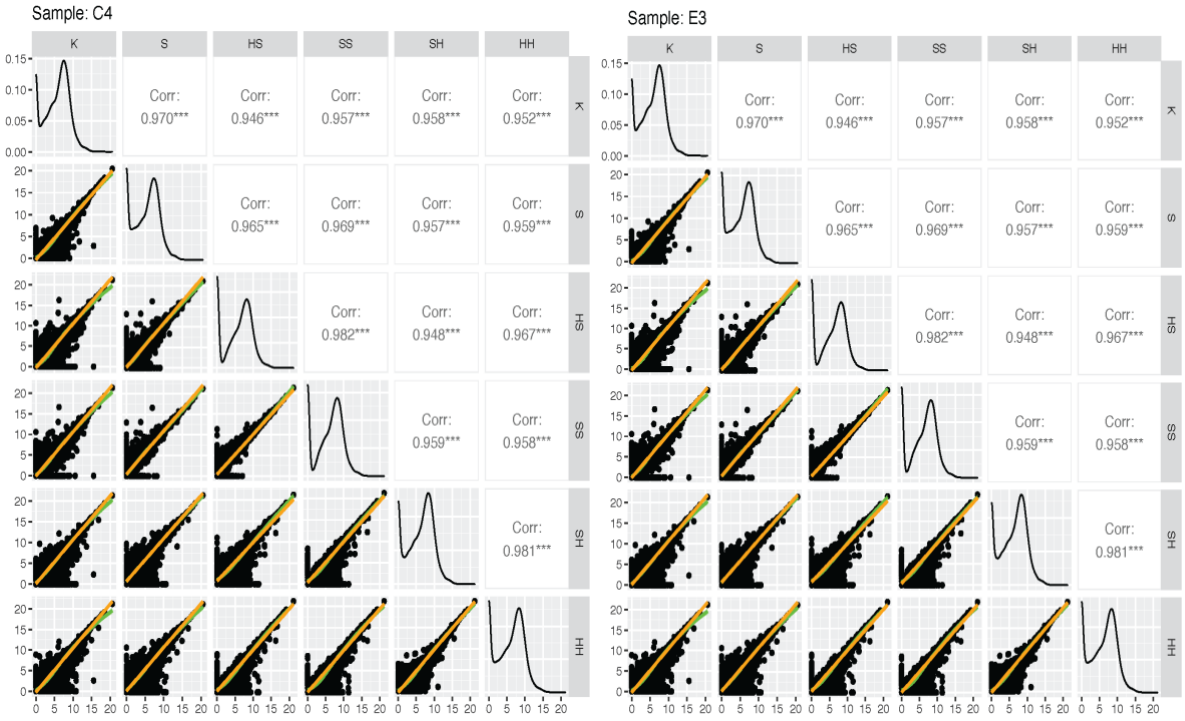

Figure 2.3: Raw count distributions (diagonal) and correlations across quantification methods for 2 samples.

Scatter matrices for transformed raw counts ($\log_2$(raw counts +1)) for two samples (one from each treatment) to visualize count distributions generated by these tools and correlation metrics. K and S refer to the pseudoaligners Kallisto and Salmon. HS and HH refer to Hisat2 aligner and either StringTie (HS) or HTSeq (HH) counters. SS and SH refer to STAR aligner and either StringTie (SS) or HTSeq (SH) counters. The top triangle of the matrix are Spearman correlation estimates. *** indicate p-values < 0.0001. Orange lines are loess fits to the scatter matrices in the lower triangle.

These results were extended to a more quantitative comparison by calculating R_v coefficients for count matrices between the 6 aligner/counter combinations. Rv coefficients are metrics of similarity for matrices, where values of 0 indicates two matrices are completely different and values of 1 indicate two matrices are the exact same (Smilde *et al.* 2009). The Rv coefficients for the unfiltered data demonstrate high similarity in our comparison across pipelines (aligner/counter combinations) with values ranging from 0.907 – 0.977 (Figure 2.4). The high level of similarity, and the individual relationships discussed in the scatter matrices are well-supported by these analyses and extend relationships identified across all samples.

Differential Gene Expression

Count matrices from all 6 aligner/counter combinations were run in DESeq2, EdgeR, and Limma-Voom for DGE detection in a comparison between caloric restricted and ad lib treatment groups. Across the three approaches of pre-filtering low and no expression genes, we see the same general pattern in the percent of statistically significant DEGs (no Fold Change cut-off) detected for each pipeline. DESeq2 always detects the highest amount (30.94 – 40.88%) of DEGs, followed by Limma-Voom (24.75 – 30.17%), and lastly, the most stringent detection method, EdgeR (11.69 – 16.28%). We added an additional filter for biologically significant changes in expression values by filtering for DEGs with a $\log_2$ fold change of the absolute value of 2 or more. These data were used in the remaining analyses and are referred to as significant DEGs.

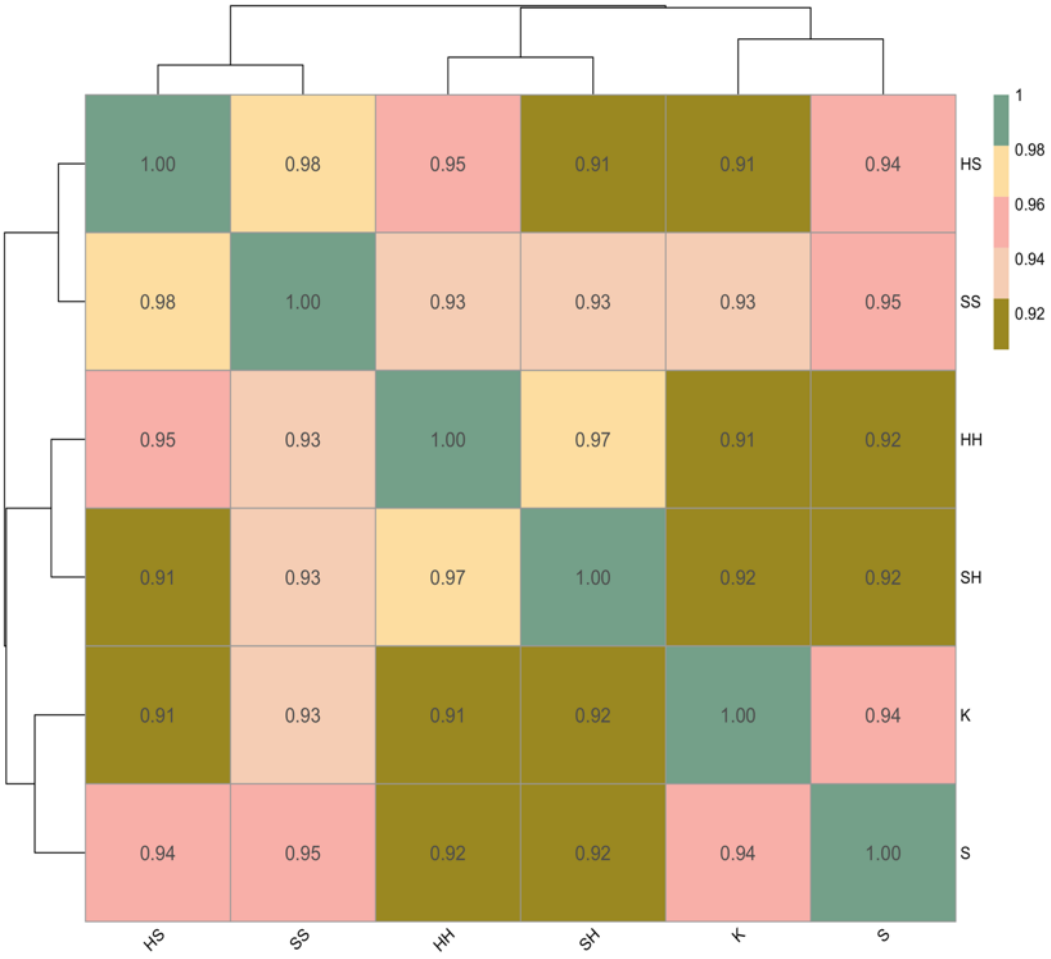

Figure 2.4: High similarity between quantification methods across samples based on R_v coefficients.

A heatmap of R_v coefficients indicating the level of similarity between raw count matrices across samples. Colors correspond to the color legend on the top right, where 1 indicates the matrices are the same, and zero (not shown) indicates matrices are completely different. K and S refer to the pseudoaligners Kallisto and Salmon, HS and HH refer to Hisat2 aligner and either StringTie (HS) or HTSeq (HH) counters, SS and SH refer to STAR aligner and either StringTie (SS) or HTSeq (SH) counters. P-values of zero from 1000 permutations were reported for these values.

We used a general linear model (lm) to test the relative effects of each step on the number of significant DEGs. Although our response variable was count data, a lm provided the best fit for our data compared to other count-based models.

Model 1: Effects of quantification and DGE programs on number of significant DEGs

Data for model 1 was analyzed separately for each pre-filtering method to avoid committing pseudoreplication with data that naturally precedes filtering steps (i.e., quantification). Here, we summarize across models, but predominantly present figures and tables from pipeline-specific filtering results in the main text and other methods in the supplementary material (Appendix 1). A majority of the total variance (total sum of squares, TSS) is explained by DGE program (73 – 83%; F 78.46 – 271.28; $p < 0.001$), with another 15 - 22% (F 9.40 – 29.94; $p < 0.001$) being explained by quantification method (aligner/counter) and the remaining 2 or 5% residual error (Table2.2).

Focusing on data generated with pipeline-specific filtering, we see the largest (statistically significant) deviations from the grand mean due to quantification method includes (i) Hisat2 or STAR with HTSeq pipelines exceeding the mean, and (ii) Hisat2 with StringTie pipelines being lower than the grand mean. Sum coded model summaries with these results are displayed in Table 2.3 and plotted in Figure 2.5 and should be interpreted as deviations from the grand mean of filtered DEG (intercept). This coding seemed more intuitive to interpret the effects of the predictors, than comparing each level of a predictor to the alphanumeric reference for that predictor. DESeq2 and EdgeR pipelines had 76 and 21 fewer DEG than the grand mean, respectively. Limma-Voom pipelines exceeded the grand mean with an average of 97 more DEG (not shown). Although DESeq2 had the highest amount of DEG when filtering for statistical

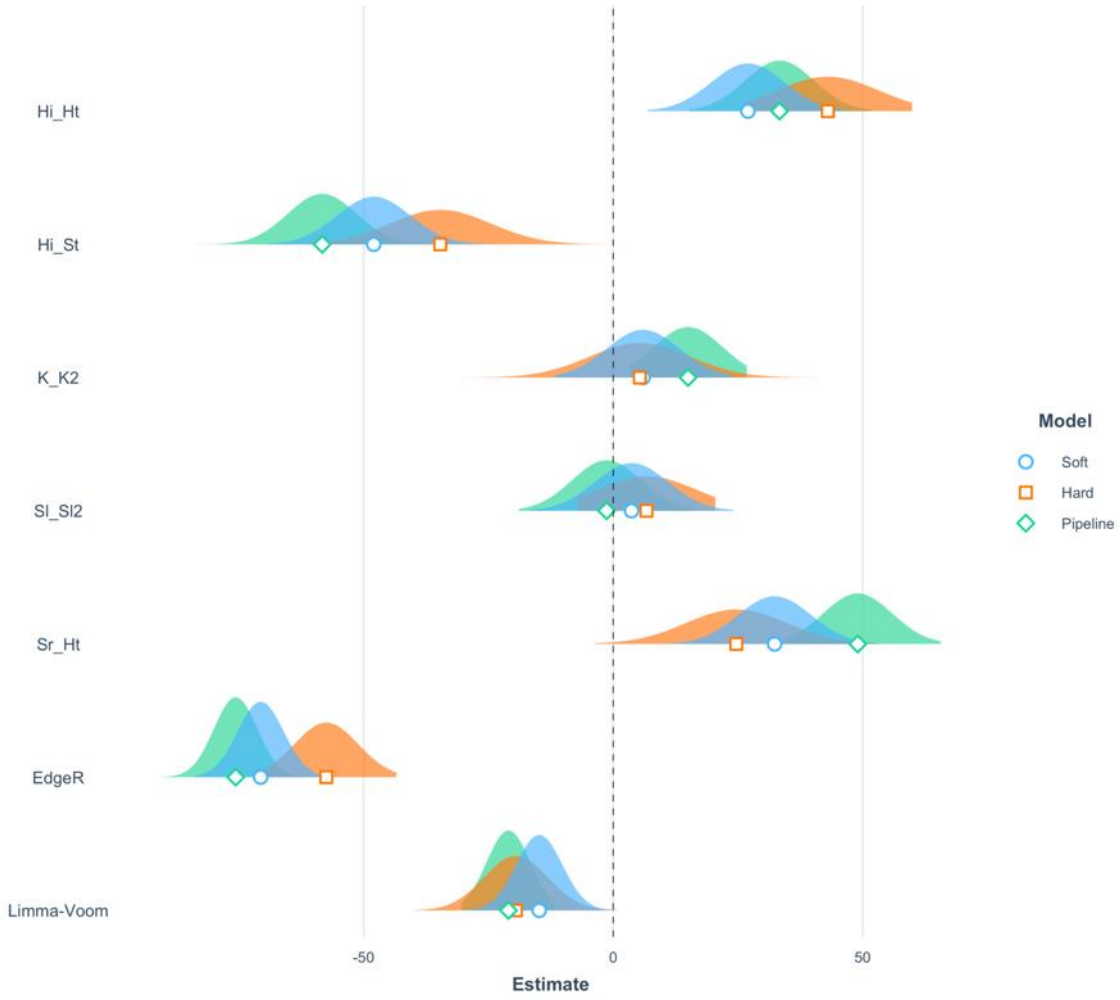

Figure 2.5. Plotted linear model 1 summaries from Table 2.3.

Model estimates are on the x-axis for three models (one from each pre-filtering approach). Shapes are coefficient estimates for predictor variables (quantification and DGE programs) interpreted as average differences from the grand mean, where values close to zero are close to the grand mean. Normal distributions represent theoretical values around the mean estimates for each program or combination of programs to visualize uncertainty. Hi & Sr are aligners Hisat2 and STAR. Ht & St are counters HTSeq and StringTie. K & S are pseudoaligners Kallisto and Salmon; The added number 2 were for coding purposes to indicate alignment and counting steps were carried out by the same program.

significance only, adding a filter for differentially expressed genes with large changes puts this pipeline in last place. In proportion to the larger number of significant DEGs detected, DESeq2 has the lowest amount of DEGs with large fold changes (~ 1/5 of DEGs at an FDR of 5%), which could be effect of this program's dispersion estimation and shrinkage methods, relative to EdgeR and Limma-Voom (Soneson and Delorenzi 2013). Lastly, pipelines with pseudomappers did not have statistically significant deviations from the mean (Salmon -1.33 ; Kallisto 15; SE 6.90). Post hoc analysis with emmeans identified 112 statistically significant (Tukey's adjusted) pairwise mean comparisons between combination of quantification and DGE programs that can be explored in Table S2.1 for pipeline-specific filtered data (other filter methods can be found in Tables S2.2 & S2.3). Overall patterns from these estimates demonstrate the highest mean differences in number of predicted DEGs for pipelines being contrasted with Limma-Voom and HTSeq pipelines which can be best visualized with plotted adjusted predictions in Figure 2.6. These analyses demonstrate that a combination of STAR and HTSeq with Limma-Voom for DGE detection report the most statistically significant DEGs with moderate to high fold changes. While we see a larger proportion of the variation in our sample being explained by DGE programs, we were not able to fully decompose variation for both quantification steps (alignment and counting) and didn't have enough degrees of freedom to test for interactions with these data. Our second model was generated to investigate this relationship and test for interactions between our predictor variables.

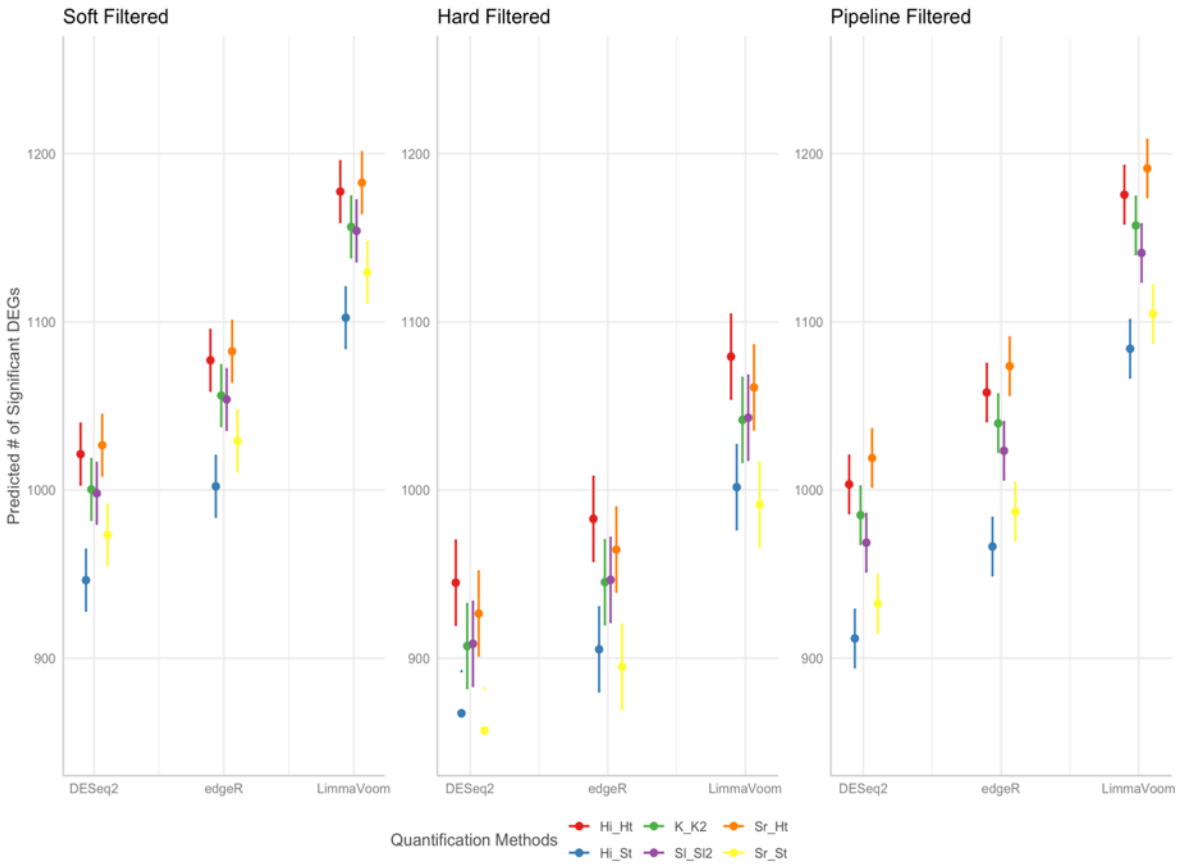

Figure 2.6: HTSeq and Limma-Voom increase the number of significant DEGs detected across pre-filtering methods.

Plotted predicted (estimated marginal) means for Model 1 faceted by DGE detection method. Coefficient estimates from Figure 2.5 added to the grand mean produce predicted mean number of DEGs for each combination of RNA-seq analysis programs. Lines extending from points represent confidence intervals around the estimates. Results are paneled by pre-filtering approaches. See Table 2.1 for program abbreviations; the added number 2 were for coding purposes to indicate alignment and counting steps were carried out by the same program.

Model 2: Effects of aligner, counter, and DGE programs on filtered DEGs

We exclude pseudoaligner pipelines from model 2 data, as discussed in the methodology. Filtering methods were also analyzed separately for these data, and as with previous results, we discuss broad summaries about all three methods and pipeline-specific filtering results for relevance and applicability. Across pre-filtering methods, we find that the variation due to quantification steps is largely attributed to the choice of counter (18.6 – 31.41%; F 157.79 – 4349.92; $p < 0.001$), while aligner explains ~ 1% of the variation in our sample (F 10.02 – 180.56; $p < 0.025$ or less). The ANOVA (Table S2.2) recapitulated the importance of DGE program (62.99 – 77.5%; F 328.48 – 6117.31; $p < 0.001$) and reported a significant interaction between counter and DGE programs accounting for 1.5 – 4% of the total variation (F 8.95 – 125.12; $p < 0.025$). Model summaries visualized in Figure 2.7 (see Table S2.3 for tabular output) recapitulate relationships in our previous model for DGE programs where DESeq2 and EdgeR pipelines had significant negative deviations from the overall mean and Limma-Voom pipelines exceeding that of the mean across pipelines. Aligner choices cause an average mean deviation of $^{+}/_{-}$ 9 genes for STAR and Hisat2 pipelines, respectively. Counter choice causes deviations from the mean by a magnitude of ~ 45, meaning the mean of HTSeq pipelines exceed the grand mean by 45 DEGs. The positive deviation from the mean when combining HTSeq with DESeq2 was statistically significant, but the mean deviation for combinations of HTSeq with Limma-Voom detecting ~133 more DEGs than the overall mean was the largest deviation observed. These linear models provide insight for how the number of significant genes with moderate to large expression change

is influenced by different steps of DGE analysis, but it would be helpful to also look at the overlap of these list to understand difference in DEG content.

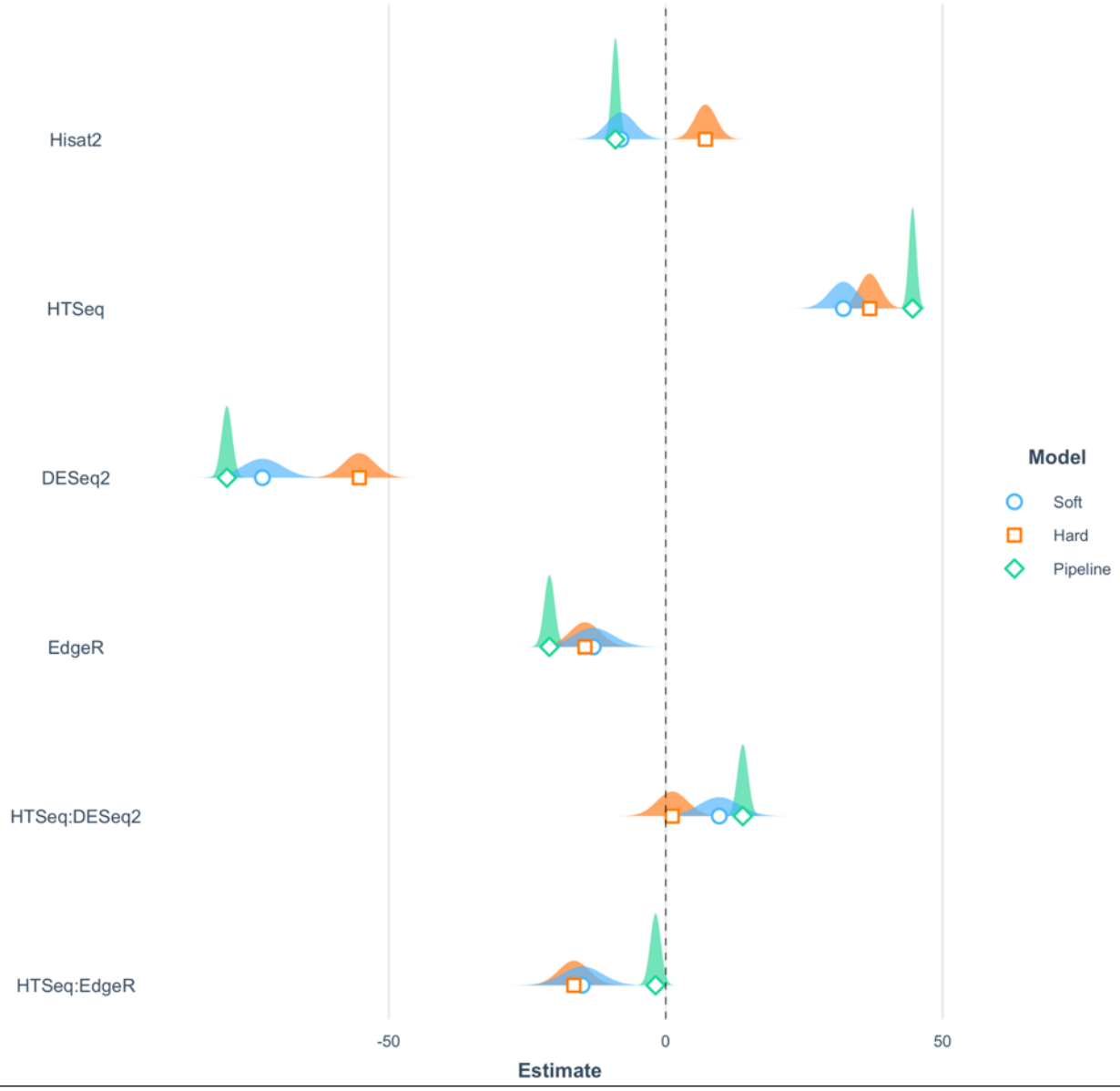

Figure 2.7. Plotted linear model 2 summaries from Table S2.3

Model estimates are on the x-axis for three models (one from each pre-filtering approach). Shapes are coefficient estimates for predictor variables (quantification and DGE programs) interpreted as average differences from the grand mean, where values close to zero are close to the grand mean. Normal distributions represent theoretical values around the mean estimates for each program or combination of programs. Hi & Sr are aligners Hisat2 and STAR. Ht & St are counters HTSeq and StringTie. K & S are pseudoaligners Kallisto and Salmon; The added number 2 were for coding purposes to indicate alignment and counting steps were carried out by the same program.

The highest levels of similarity was seen between pipelines containing EdgeR or DESeq2, particularly when the same quantification steps were used. Limma-Voom pipelines had moderate to high levels of similarity with each other and lower levels with other DGE programs, but this could be an artifact of large differences in gene set sizes for other pipelines relative to Limma-Voom. Upset plots show~85% of the possible overlap (size of the smallest set in the comparisons) is shared within DESeq2 and EdgeR pipelines and 84% between them. Limma-Voom had just as much overlap within pipelines as there was between pipelines using the other two methods. The higher levels of variation within Limma-Voom pipelines could be an indication of the robustness of this model across quantification methods, particularly the interaction with the counter algorithm preceding it. These visualizations also reflect our model where pipelines with Limma-Voom consistently produce more filtered DEGs relative to the mean (gene set size in Upset plots – Figure 2.8; Tables 2.3 & S2.3). DESeq2 and EdgeR pipelines both shared a 71% overlap with Limma-Voom pipelines with 172 or 148 Limma-Voom specific genes, respectively. In DESeq2 and EdgeR comparisons, unique genes were few (12 in EdgeR) to non-existent, further highlighting the similarity between these two methods. These relationships are maintained across filtering methods (Figure 2.8; Figures S2.5 – S2.6).

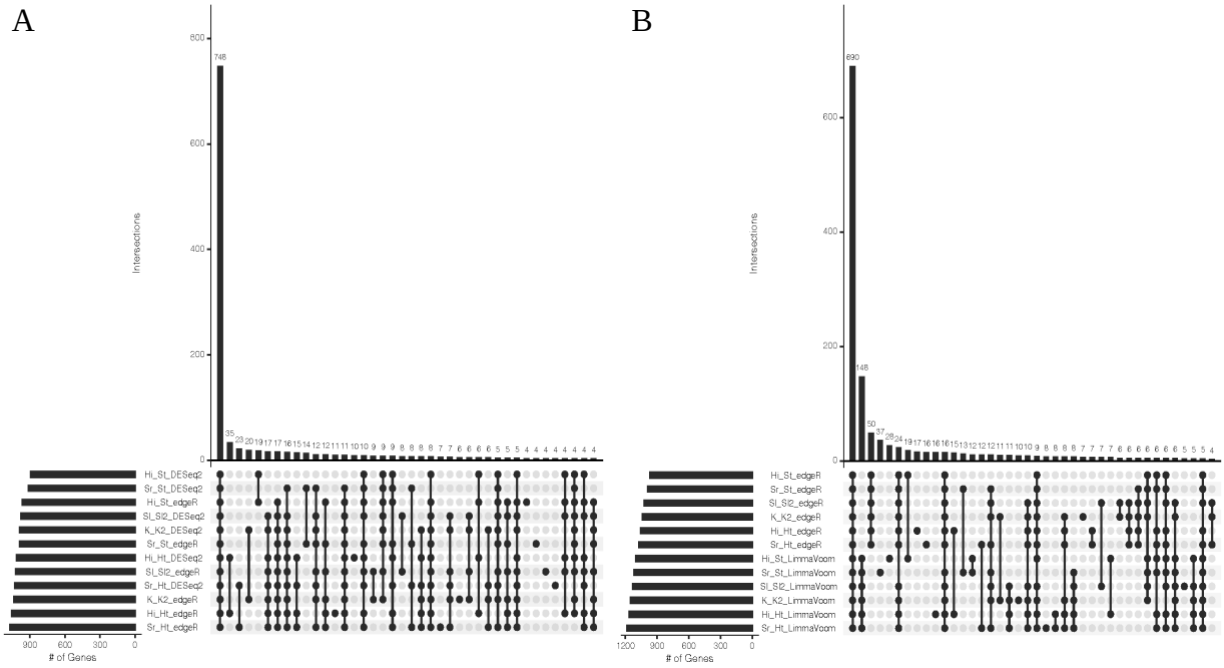

Figure 2.8. High overlap between DGE programs, particularly those using Negative Binomial Models

Upset plots of DEGs for pipeline-specific filtering. DEG intersections for pipelines including (A) EdgeR or DESeq2 and (B) EdgeR or Limma-Voom. Horizontal bars represent the total number of DEGs for a pipeline. Filled circles indicate the specific pipelines being intersected and the vertical bars are the number of DEGs in that intersection.

Together, these analyses explore the variation in the number of significant, moderate to large effect DEGs (fold change of 4; FDR 5%) due to the analysis steps used to produce them. Overall, we find high similarity among quantification methods with similarity estimates at 91% or higher across methods. This agrees with previous work done in Schaarschmidt et al. (2021) who found high levels of similarity across 7 aligners and in Corchete et al. (2021) (Schaarschmidt *et al.* 2020; Corchete *et al.* 2020). The effects of these programs are dependent upon the type of RNA-analysis being performed. Work done by Wu et al. (2018) aligns with the relationships identified here and in other comparisons, but only in longer RNA molecules. Kallisto and Salmon performed poorly for lowly expressed and small genes due to SNPs (Wu *et al.* 2018). This work is one of many that highlights the importance of context in expression analysis as performance results may not hold for all types of RNA-seq analyses or at all levels of expression. Our findings were all in the context of gene-level expression analyses, and seemingly do not agree with Srivastava et al. (2020) who found more variability between traditional aligners and alignment free methods in transcript-level expression analyses (Srivastava *et al.* 2020). DGE and counting programs were the largest sources of variation and they interacted strongly with each other. The mean number of DEG for Limma-Voom pipelines had the most extensive differences, greatly exceeding the mean across all pipelines, and had the lowest levels of similarity when compared to pipelines with DESeq2 or EdgeR. In contrast, DESeq2 and EdgeR pipelines were usually below the overall mean, but these differences were lower for EdgeR and were much closer to the overall mean. The high levels of similarity between DESeq2 and EdgeR were quite clear

across data analyses, corroborating previous findings, due to the shared underlying count model between these methods (Soneson and Delorenzi 2013; Seyednasrollah *et al.* 2015).

Normalization, parameter shrinking, and count modeling methods are the main components of DGE programs. These steps are performed similarly in DESeq2 and EdgeR, but very differently in Limma-Voom. Limma-Voom uses a log-normal distribution to model counts and makes calculations using geometric means, while the other two algorithms use a negative binomial distribution to model counts and performs calculations with raw (normalized) counts (Robinson *et al.* 2010; Law *et al.* 2014; Love *et al.* 2014). Negative bionomial models are generally better suited for modeling count data and are easier to interpret contextually, but often come with higher Type I error rates. Many statisticians advise researchers to prioritize the model that best fit the data and worry about things that can be corrected (i.e., false positives) secondarily. Yet they still circle back to the fact that all models are incorrect and linear models have their benefits (i.e., robust, good false positive control) particularly in statistically complex model (Warton *et al.* 2016). While it is pertinent to understand the effects of the underlying algorithms used, it is important to discuss these differences in terms of the interpretations made from them. Although there were significant differences in means for many pipeline combinations, exploring DEG list content across pipelines revealed a minimum of 71% overlap (Limma-Voom with other DGE programs) and a maximum of 84% overlap (EdgeR with DESeq2). The list contents were largely shared across all pipelines intersected, suggesting that overall differences of analysis steps are mitigated and balance out to produce highly similar DEG lists. Studies that could estimate precision and accuracy report Limma-Voom DEGs in high agreement qRT-PCR

expression analyses. This contrasts with DESeq2 and the most basic EdgeR model, which fall close to the bottom of the ranking (Corchete *et al.* 2020). While we do not explore these metrics here, we do find the largest set of hypotheses generating DEGs with Limma-Voom pipelines. Next, we explore how the variation in filtered DEGs from our pipelines influence biological conclusions drawn from gene set enrichment analyses.

Pathway Analysis

After DGE analysis, we analyzed our 18 pipelines in the R package fgsea for functional gene set enrichment analysis. Across pre-filtering methods, over half of the pipelines (78%) have no statistically significant enriched gene sets. While we have previously reported results primarily in the context of pipeline-specific filtered data, we will focus on hard filtered data. Of the 16,612 genes in our annotation, only ~4,000 gene names map the gene IDs, which is necessary for gene set enrichment analyses. FDR was set at 25%, as suggested by the GSEA documentation for hypothesis generating exploration of the data set. The pipelines that report a statistically significant enriched gene set are those using EdgeR or Limma-Voom for DGE detection. All report the same result hallmark gene set as being overrepresented by genes upregulated in the food restricted group, Xenobiotic Metabolism.

Xenobiotic metabolism refers to the detection and breakdown of exogenous chemicals (i.e., plant compounds, drug, cosmetics) that may or may not be considered toxins but are metabolized and excreted from the body (Johnson *et al.* 2012). Interestingly, the same nuclear receptors that respond to xenobiotic compounds also respond to endobiotic compounds, particularly lipids. This has been documented in

humans, flies, and several other arthropods including our organism of interest, *D. pulex*. Specifically, HR96, a xeno- and endobiotic nuclear receptor, and its orthologs are involved in toxicant response and cholesterol homeostasis. Unsaturated fatty acids are also common regulators (mostly activators) of receptor genes in xenobiotic metabolism. Xenobiotic metabolism genes also regulate stress responses including responses to starvation (Karimullina *et al.* 2012). *Daphnia* tend to maintain higher levels of unsaturated fatty acids relative to the content in their diet and will sequester these lipids when they are starving, or their food quality is poor (Brett *et al.* 2006). This is a confirmatory result of the caloric restriction treatment and may indicate that *Daphnia* metabolize stored unsaturated fatty acids during periods of food restriction or poor diet. A quick look at the leading-edge genes supporting this pathway as enriched include genes related to transcription factors involve in lipid metabolism and toxicity (ABCD2) and genes that are recognized from cellular metabolism pathways (FBP1, IDH1, MTHFD1). The lipid metabolism pathways are documented to be active during periods of acute starvation after carbohydrate stores are depleted (Campos *et al.* 2021). Lipids are more slowly metabolized, balanced by the metabolism of intermediates such as glycerol (from lipolysis) and acetyl-CoA (from glycolysis) for energy generation (Klumpen *et al.* 2021). These mechanisms are less understood in the context of chronic starvation stress induced in our experiment. These results should be interpreted gingerly because 75% of gene IDs in our annotation lack gene names and restricts GSEA analyses to a much smaller gene subset to make interpretations from.

FUTURE DIRECTIONS & CONCLUSIONS

Annotation quality and completeness is extremely important in RNA-seq analyses to infer the biological pathways that are being affected by treatments, as we can see from the limitations in GSEA applied to these data. This resource is also important to the overall DGE analysis procedures because the annotation is responsible for identifying features that reads are mapping to. This highlights the need for the ecological and evolutionary biology communities to "take a page from the book of model organism research" and begin building infrastructure for non-model omics resources widely available to the community that represents the biodiversity we love and explore in our research. In future analyses, we will explore the consensus of biological interpretation across DGE pipelines in different non-model data sets with more complete, higher quality genome annotations used by our lab group (e.g., garter snake, fence lizard (Westfall *et al.* 2021)). We use the default models and procedures for these comparisons, but the DGE programs we applied all have different normalization and modeling methods that have been added to accommodate other RNA-seq data profiles. Many of these modifications do require the user to have a strong understanding of the underlying models, the parameters, and how to assess model fit, which is not an inherent skillset for many biologist and budding researchers interested in applying these methods in their investigations. As more RNA-seq data is produced across more systems, we expect that analysis algorithms will develop that are more robust and generalizable across experimental designs and species.

Our analyses were designed from the perspective of biologists with limited experience with statistical modeling and omics data to understand how the decisions (i.e.,

analysis program choice, filtering methods) and resources (i.e., annotation) involved in DGE analysis change the overall biological interpretations that are driving the investigations. We recapitulate findings from previous researchers on the relatively miniscule importance of mapping strategy, moderate importance of counter choice, and high importance of normalization and count modeling methods in DGE analysis. Yet, we find that even considering these effects, and the caveats of our data, we recover the same pathways across pipelines (even ones that lack statistical significance) with a treatment relevant pathway at the top of these lists, suggesting the specific choice of the programs at any one of these steps may not matter as much in the context of the biological interpretations of the experiment. We did not test precision and accuracy for these data, so our conclusions are not in the context of the best model and pipelines to use. We do conclude that researchers interested in obtaining the maximum number of significant DEGs should use a combination of STAR for alignment, HTSeq for counting, and Limma-Voom for DGE analysis. We would like to remind individuals that these pipelines are not "black box" solutions to RNA-seq analyses but are starting places to learn how to perform classic control-treatment RNA-seq and DGE analyses. Users are advised to explore other settings, parameters, and programs that are best for their data set.

Finally, we aim for the code generated to test these pipelines to be useful for instructors to teach upper-level undergraduate and graduate students to perform RNA-seq analysis. We intend for this code to allow users to focus more on the conceptual background and biological interpretation for these types of data and less on how to script it. We make this code available on GitHub in a modular format with the appropriate

background information and helpful online resources to enrich learning. These code products are useful as a self-paced tutorial, or modules incorporated into genetics or functional genomics curriculum. As an example, these resources will be integrated into the Functional Genomics course at Auburn University in the format of a course-based research experience (CRE) that allow students a hands-on learning experience with a computationally tractable, publicly available dataset. Students will have the opportunity to contrast methodologies, while working together to complete RNA-seq analysis from alignment to biological interpretation. These analyses are relatively fast and can be easily used with other publicly available datasets with control-treatment designs. We hope that the computational biology and bioinformatics learning community and other readers find these resources beneficial and informative.

Data Availability

All sequence data are available under the NCBI BioProject Accession PRJNA437447. Code used to perform the data analyses for this work can be found on GitHub (https://github.com/Schwartz-Lab-at-Auburn/18_RNA-seq_Pipelines).

Acknowledgements

We acknowledge the students in the 2016, 2019, and 2021 Functional Genomics course for their patience as we test out these early teaching methodologies. We give special acknowledgement to the following students and collaborators: Seungyeon Seo, Mursalin Khan, Breanna Sipley, Ambika Pokhrel, Rory Telemeco, Damien Waites, Chidozie Ugochukwu, Stephen Tansie. We acknowledge Julia Golke for providing and

environment and resources for conducting the caloric restriction experiment, and Phil Pearson for assistance with the experiment.

Funding

James S. McDonnell Foundation Postdoctoral Fellowship (TSS) provided resources and support for TSS. Support to TSS provided by Office of Energetics during data collection. NIH-1R15AG064655-01 to TSS supported AC during the analysis of these data.

REFERENCES

Abrams, Z. B., T. S. Johnson, K. Huang, P. R. O. Payne, and K. Coombes, 2019 A protocol to evaluate RNA sequencing normalization methods. BMC Bioinformatics 20: 679.

Anders, S., P. T. Pyl, and W. Huber, 2015 HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

Anderson, R. M., and R. Weindruch, 2010 Metabolic reprogramming, caloric restriction and aging. Trends Endocrinol. Metab. TEM 21: 134–141.

Andrews, Simon, 2010 FASTQC. A quality control tool for high throughput sequence data.

Asselman, J., J. D. Hochmuth, and K. A. C. De Schamphelaere, 2014 A comparison of the sensitivities of *Daphnia magna* and *Daphnia pulex* to six different cyanobacteria. Harmful Algae 39: 1–7.

Asselman, J., M. E. Pfrender, J. A. Lopez, J. R. Shaw, and K. A. C. De Schamphelaere, 2018 Gene coexpression networks drive and predict reproductive effects in *Daphnia* in response to environmental disturbances. Environ. Sci. Technol. 52: 317–326.

Baruzzo, G., K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald *et al.*, 2017 Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods 14: 135–139.

Becker, D., Y. Reydelet, J. A. Lopez, C. Jackson, J. K. Colbourne *et al.*, 2018 The transcriptomic and proteomic responses of *Daphnia pulex* to changes in

temperature and food supply comprise environment-specific and clone-specific elements. BMC Genomics 19: 376.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bourgon, R., R. Gentleman, and W. Huber, 2010 Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. 107: 9546–9551.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter, 2016 Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34: 525–527.

Brett, M. T., D. C. Müller-Navarra, A. P. Ballantyne, J. L. Ravet, and C. R. Goldman, 2006 Daphnia fatty acid composition reflects that of their diet. Limnol. Oceanogr. 51: 2428–2437.

Brunner, F. S., J. A. Deere, M. Egas, C. Eizaguirre, and J. A. M. Raeymaekers, 2019 The diversity of eco-evolutionary dynamics: Comparing the feedbacks between ecology and evolution across scales. Funct. Ecol. 33: 7–12.

Campos, B., B. Piña, and C. Barata, 2021 *Daphnia* magna Gut-Specific Transcriptomic Responses to Feeding Inhibiting Chemicals and Food Limitation. Environ. Toxicol. Chem. 40: 2510–2520.

Chislock, M. F., R. B. Kaul, K. A. Durham, O. Sarnelle, and A. E. Wilson, 2019a Eutrophication mediates rapid clonal evolution in *Daphnia pulicaria*. Freshw. Biol. 64: 1275–1283.

Chislock, M. F., O. Sarnelle, L. M. Jernigan, V. R. Anderson, A. Abebe *et al.*, 2019b Consumer adaptation mediates top–down regulation across a productivity gradient. Oecologia 190: 195–205.

Colbourne, J. K., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker *et al.*, 2011a The ecoresponsive genome of *Daphnia pulex*. Science 331: 555–561.

Colbourne, J. K., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker *et al.*, 2011b The Ecoresponsive Genome of *Daphnia* pulex. Science 331: 555–561.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera *et al.*, 2016 A survey of best practices for RNA-seq data analysis. Genome Biol. 17: 13.

Corchete, L. A., E. A. Rojas, D. Alonso-López, J. De Las Rivas, N. C. Gutiérrez *et al.*, 2020 Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci. Rep. 10: 19737.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Eads, B. D., J. Andrews, and J. K. Colbourne, 2008 Ecological genomics in *Daphnia*: stress responses and environmental sex determination. Heredity 100: 184–190.

Ebert, D., 2005 *Introduction to Daphnia Biology*. National Center for Biotechnology Information (US).

Engström, P. G., T. Steijger, B. Sipos, G. R. Grant, A. Kahles *et al.*, 2013 Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods 10: 1185–1191.

Freshney, R. I., 2005 Quantitation.

Hairston, N. G., C. L. Holtmeier, W. Lampert, L. J. Weider, D. M. Post *et al.*, 2001 Natural selection for grazer resistance to toxic cyanobacteria: Evolution of phenotypic plasticity? Evolution 55: 2203–2214.

Jia, J., G. Hu, C. Feng, C. Dong, M. Han *et al.*, 2020 *Daphnia carinata* genome provides insights into reproductive switching: Preprints preprint.

Johnson, C. H., A. D. Patterson, J. R. Idle, and F. J. Gonzalez, 2012 Xenobiotic Metabolomics: Major Impact on the Metabolome. Annu. Rev. Pharmacol. Toxicol. 52: 37–56.

Jones, O. R., A. Scheuerlein, R. Salguero-Gómez, C. G. Camarda, R. Schaible *et al.*, 2014 Diversity of ageing across the tree of life. Nature 505: 169–173.

Josse, J., and S. Holmes, 2016 Measuring multivariate association and beyond. Stat. Surv. 10: 132–167.

Karimullina, E., Y. Li, G. Ginjupalli, and W. S. Baldwin, 2012 *Daphnia* HR96 is a Promiscuous Xenobiotic and Endobiotic Nuclear Receptor. Aquat. Toxicol. Amst. Neth. 116–117: 69–78.

Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37: 907–915.

Klumpen, E., N. Hoffschröer, A. Schwalb, U. Gigengack, M. Koch *et al.*, 2021 Metabolic adjustments during starvation in *Daphnia* pulex. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 255: 110591.

Koch, C. M., S. F. Chiu, M. Akbarpour, A. Bharat, K. M. Ridge *et al.*, 2018 A Beginner's Guide to Analysis of RNA Sequencing Data. Am. J. Respir. Cell Mol. Biol. 59: 145–157.

Korotkevich, G., V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov *et al.*, 2021 Fast gene set enrichment analysis. 060012.

Kukurba, K. R., and S. B. Montgomery, 2015 RNA Sequencing and Analysis. Cold Spring Harb. Protoc. 2015: pdb.top084970.

Lack, J. B., L. J. Weider, and P. D. Jeyasingh, 2018 Whole genome amplification and sequencing of a *Daphnia* resting egg. Mol. Ecol. Resour. 18: 118–127.

Law, C. W., Y. Chen, W. Shi, and G. K. Smyth, 2014 voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15: R29.

Lee, B.-Y., B.-S. Choi, M.-S. Kim, J. C. Park, C.-B. Jeong *et al.*, 2019 The genome of the freshwater water flea *Daphnia magna*: A potential use for freshwater molecular ecotoxicology. Aquat. Toxicol. 210: 69–84.

Liberzon, A., C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov *et al.*, 2015 The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 1: 417–425.

Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

Markowetz, F., 2017 All biology is computational biology. PLOS Biol. 15: e2002050.

Miner, B. E., L. De Meester, M. E. Pfrender, W. Lampert, and N. G. Hairston, 2012 Linking genes to communities and ecosystems: *Daphnia* as an ecogenomic model. Proc. R. Soc. B Biol. Sci. 279: 1873–1882.

Musich, R., L. Cadle-Davidson, and M. V. Osier, 2021 Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Front. Plant Sci. 12:.

Nelson, J. R., T. S. Schwartz, and J. M. Gohlke, 2018 Influence of maternal age on the effects of seleno-l-methionine in the model organism *Daphnia pulex* under standard and heat stress conditions. Reprod. Toxicol. 75: 1–9.

Nickel, J., T. Schell, T. Holtzem, A. Thielsch, S. R. Dennis *et al.*, 2021 Hybridization dynamics and extensive introgression in the *Daphnia longispina* species complex: new insights from a high-quality *Daphnia galeata* reference genome. bioRxiv 2021.02.01.429177.

Nobel Media AB 2021 Ilya Mechnikov – Facts. NobelPrize.org.

Osborne, T. B., L. B. Mendel, and E. L. Ferry, 1917 The Effect of Retardation of Growth Upon the Breeding Period and Duration of Life of Rats. Science 45: 294–295.

Paerl, H. W., R. S. Fulton, P. H. Moisander, and J. Dyble, 2001 Harmful freshwater algal blooms, with an emphasis on cyanobacteria. Sci. World J. 1: 76–113.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, 2017 Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14: 417–419.

Pelletier, F., D. Garant, and A. P. Hendry, 2009 Eco-evolutionary dynamics. Philos. Trans. R. Soc. B Biol. Sci. 364: 1483–1489.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33: 290–295.

Plaisier, S. B., R. Taschereau, J. A. Wong, and T. G. Graeber, 2010 Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. Nucleic Acids Res. 38: e169.

Qiagen QIAGEN CLC Genomics Workbench. Bioinforma. Softw. Serv. QIAGEN Digit. Insights.

Reimand, J., R. Isserlin, V. Voisin, M. Kucera, C. Tannus-Lopes *et al.*, 2019 Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat. Protoc. 14: 482–517.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law *et al.*, 2015 limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43: e47.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. 26: 139–140.

Sarnelle, O., S. Gustafsson, and L.-A. Hansson, 2010 Effects of cyanobacteria on fitness components of the herbivore *Daphnia*. J. Plankton Res. 32: 471–477.

Sarnelle, O., and A. E. Wilson, 2005 Local adaptation of *Daphnia pulicaria* to toxic cyanobacteria. Limnol. Oceanogr. 50: 1565–1570.

Schaarschmidt, S., A. Fischer, E. Zuther, and D. K. Hincha, 2020 Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana. Int. J. Mol. Sci. 21: 1720.

Seyednasrollah, F., A. Laiho, and L. L. Elo, 2015 Comparison of software packages for detecting differential expression in RNA-seq studies. Brief. Bioinform. 16: 59–70.

Shaw, J. R., J. K. Colbourne, J. C. Davey, S. P. Glaholt, T. H. Hampton *et al.*, 2007 Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics 8: 477.

Smilde, A. K., H. A. L. Kiers, S. Bijlsma, C. M. Rubingh, and M. J. van Erk, 2009 Matrix correlations for high-dimensional data: the modified RV-coefficient. Bioinformatics 25: 401–405.

Soneson, C., and M. Delorenzi, 2013 A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14: 91.

Soneson, C., M. I. Love, and M. D. Robinson, 2016 Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.

SRA Toolkit Development Team NCBI SRA Toolkit. GitHub.

Srivastava, A., L. Malik, H. Sarkar, M. Zakeri, F. Almodaresi *et al.*, 2020 Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. 21: 239.

Stark, R., M. Grzelak, and J. Hadfield, 2019 RNA sequencing: the teenage years. Nat. Rev. Genet. 20: 631–656.

Warton, D. I., M. Lyons, J. Stoklosa, and A. R. Ives, 2016 Three points to consider when choosing a LM or GLM test for count data. Methods Ecol. Evol. 7: 882–890.

Westfall, A. K., R. S. Telemeco, M. B. Grizante, D. S. Waits, A. D. Clark *et al.*, 2021 A chromosome-level genome assembly for the eastern fence lizard (Sceloporus undulatus), a reptile model for physiological and evolutionary ecology. GigaScience 10: giab066.

Wilson, A. E., and M. F. Chislock, 2013 Ecological control of cyanobacterial blooms in freshwater ecosystems, pp. 213–221 in *Cyanobacteria: Ecology, Toxicology and Management*, edited by Ferrão-Filho, Aloysio Da S. Nova Science Publishers, Inc, New York, NY.

Wu, D. C., J. Yao, K. S. Ho, A. M. Lambowitz, and C. O. Wilke, 2018 Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics 19: 510.

Ye, Z., S. Xu, K. Spitze, J. Asselman, X. Jiang *et al.*, 2017 A new reference genome assembly for the microcrustacean *Daphnia pulex*. G3 GenesGenomesGenetics 7: 1405–1416.

Table 2.1: Pipeline Programs – Descriptions and Parameters Used Programs used throughout these analyses with their abbreviations when applicable, a description provided by the program, the version used, and any parameters that deviated from the default/required parameters.

Program	Abbreviation	Description (Author Sourced)	Version	Parameters (Deviation from Defaults)
Aligners				
Hisat2	Hi	"HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome. Based on an extension of BWT for graphs, we designed and implemented a graph FM index (GFM), an original approach and its first implementation."	2.2.1	--rna-strandness RF [strandeness]
STAR	Sr	"Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNA-seq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure."	2.7.5	--genomeSAindexNbases 12 [Index string size adjusted for reference genome size] --outStd SAM [output SAM to standard out] --readFilesCommand gunzip -c [read compressed input]

Counters				
HTSeq	Ht	"HTSeq is a Python package for analysis of high-throughput sequencing data. Given a file with aligned sequencing reads and a list of genomic features, a common task is to count how many reads map to each feature."	0.9.1	-s reverse [strandeness] -m intersection-nonempty [the intersection of all non-empty overlap between a gene and a unique read]
StringTie	St	"StringTie is a fast and highly efficient assembler of RNA-Seq alignments into potential transcripts. It uses a novel network flow algorithm as well as an optional de novo assembly step to assemble and quantitate full-length transcripts representing multiple splice variants for each gene locus. Its input can include not only alignments of short reads that can also be used by other transcript assemblers, but also alignments of longer sequences that have been assembled from those reads"	2.1.6	--rf [strandedness] -e [prevent novel transcripts] -G [annotation] -B [ballgown style output]
Pseudoaligners				
Kallisto	K (K2)	"A tool to quantify RNA-seq data. The kallisto algorithm uses a pseudo alignment approach to speed up the alignment procedure. The "pseudo alignment" approach can quantify	1.5.1	--rf-stranded [strandedness]

		reads without making actual alignments. Kallisto can handle paired-end and single-end reads. It reports transcripts per million mapped reads (TPM).”		
Salmon	Sl (Sl2)	“Salmon is a tool for quantifying the expression of transcripts using RNA-seq data. Salmon uses new algorithms (specifically, coupling the concept of quasi-mapping with a two-phase inference procedure) to provide accurate expression estimates very quickly (i.e. wicked-fast) and while using little memory.”	0.46.2	-l A[automatic library type detection; detect strandedness]
DGE Detection				
DESeq2		“Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on model using the negative binomial distribution.”	1.36.0	contrast = c("Treat", "Restricted", "AdLib") [set contrast] pAdjustMethod = "fdr" [use FDR correction]
EdgeR		“Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact	3.38.1	topTags(n = nrow(exacttest_output)) [get all records]

		tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce read counts, including ChIP-seq, ATAC-seq, Bisulfite-seq, SAGE and CAGE."		
Limma-Voom		"Limma is a library for the analysis of gene expression microarray data, especially the use of linear models for analysing designed experiments and the assessment of differential expression. The linear model and differential expression functions apply to all gene expression technologies, including microarrays, RNA-seq and quantitative PCR."	3.52.2	topTable(sort.by = "P", n = Inf) [get all record]
Functional Enrichment				
fgsea		"The package implements an algorithm for fast gene set enrichment analysis. Using the fast algorithm allows to make more permutations and get more fine-grained p-values, which allows to use accurate standard approaches to multiple hypothesis correction."	1.22.0	defaults

Other				
SRA Toolkit		"The SRA Toolkit and SDK from NCBI is a collection of tools and libraries for using data in the INSDC Sequence Read Archives."	2.11.0	See GitHub repository for additional details (https://github.com/Schwartz-Lab-at-Auburn/18_RNA-seq_Pipelines)
FastQC		"FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis."	0.11.9	
Trimmomatic		"Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters. These adapters can pose a real problem depending on the library preparation and downstream application."	0.39	
tximport		"Imports transcript-level abundance, estimated counts and transcript lengths, and summarizes into matrices for use with downstream gene-level analysis packages."	1.24.0	

GGally		“The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data.”	2.1.2	
FactoMineR		“Exploratory data analysis methods to summarize, visualize and describe datasets. The main principal component methods are available, those with the largest potential in terms of applications: principal component analysis (PCA) when variables are quantitative, correspondence analysis (CA) and multiple correspondence analysis (MCA) when variables are categorical, Multiple Factor Analysis when variables are structured in groups, etc. and hierarchical cluster analysis.”	2.4	
ComplexHeatmap		“Complex heatmaps are efficient to visualize associations between different sources of data sets and reveal potential patterns.”	2.12.0	
car		“An R Companion to Applied Regression.”	3.1-0	

emmeans		“Obtain estimated marginal means (EMMs) for many linear, generalized linear, and mixed models. Compute contrasts or linear functions of EMMs, trends, and comparisons of slopes. Plots and other displays.”	1.7.5	
ggeffects		“Compute marginal effects and adjusted predictions from statistical models and returns the result as tidy data frames. These data frames are ready to use with the 'ggplot2'-package. Effects and predictions can be calculated for many different models. Interaction terms, splines and polynomial terms are also supported.”	1.1.2	
UpsetR		“Creates visualizations of intersecting sets using a novel matrix design, along with visualizations of several common set, element and attribute related tasks.”	1.4.0	

Table 2.2: Analysis of Variance (ANOVA) summaries of model 1 for each filtering method. Sum of Squares values were used to calculate percent variability explained by the predictors that are reported in the text.

	Sum Sq	Df	F value	Pr (>F)
Soft Filtered				
Quantification	13706.67	5	14.24069	0.0002825
DGE Program	75144.33	2	195.18009	0.0000000
Residuals	1925.00	10		
Hard Filtered				
Quantification	17271.333	5	9.39766	0.00154
DGE Program	57693.000	2	78.47964	0.0000008
Residuals	3675.667	10		
Pipeline-Specific Filtered				
Quantification	25681.333	5	29.94326	0.0000104
DGE Program	93065.333	2	271.27478	0.0000000
Residuals	1715.333	10		

Table 2.3: Linear model 1 summaries for each filtering method. Estimates are mean differences in the number of significant DEGs from the grand mean or intercept averaged over all other predictors (i.e., Hi_Ht estimate is the average number of significant DEGs for all 3 DGE programs when quantifying with Hisat2 & HTSeq, minus the grand mean). Values below estimates in parentheses are standard error estimates. Intercept is the grand mean; see Table 2.1 for program abbreviations.

	Soft	Hard	Pipeline
(Intercept)	1065.00 ***	959.33 ***	1045.67 ***
	(3.27)	(4.52)	(3.09)
Hi_Ht	27.00 **	43.00 **	33.33 ***
	(7.31)	(10.10)	(6.90)
Hi_St	-48.00 ***	-34.67 **	-58.33 ***
	(7.31)	(10.10)	(6.90)
K_K2	6.00	5.33	15.00
	(7.31)	(10.10)	(6.90)
Sl_Sl2	3.67	6.67	-1.33
	(7.31)	(10.10)	(6.90)
Sr_Ht	32.33 **	24.67 *	49.00 ***
	(7.31)	(10.10)	(6.90)
DESeq2	-70.67 ***	-57.50 ***	-75.67 ***
	(4.62)	(6.39)	(4.37)
EdgeR	-14.83 **	-19.50 *	-21.00 ***
	(4.62)	(6.39)	(4.37)
N	18	18	18
R2	0.98	0.95	0.99

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

CHAPTER 3:

Primary culture in non-model organisms: the establishment and validation of *Anolis* lizard dermal cells

Amanda D. Clark[1], Taylor McKibben[1], Milica Courtenay[1], Aaron Reedy[2], Robert Cox[2], Tonia S. Schwartz [1]

[1]Department of Biological Sciences, Auburn University, Auburn, AL 36849

[2]Department, University of Virginia, Arlington, VA

Corresponding Author: Amanda Clark (adc0032@auburn.edu) or Tonia Schwartz (tss0019@auburn.edu)

Keywords: Reptile culture; Conservation; Biobanking

Target Journal: Molecular Ecology Resources

INTRODUCTION

Primary culture, the outgrowth of cell populations from animal tissue biopsies, has perpetually advanced basic and applied fields of biology since Harrison's demonstration of cellular outgrowth from frog embryo nerve fibers in 1907 (Keshishian 2004). Forty-five years later, this process of subculture generated the first human cell line, HeLa, was established from a cervical carcinoma sample (Landry *et al.* 2013). Fundamental molecular biology knowledge generated and deepened by cell culture methodology include, but are not limited to, studies of intercellular interaction (Saunders and D'Amore 1992), intracellular signaling (Sastry and Burridge 2000; Hartford Svoboda and Reenstra 2002), and cellular differentiation and development (Fox *et al.* 1967; Keller 2005; Myhre and Pilgrim 2010). Biomedical sciences have also made significant strides by applying cell culture techniques in pharmacology and toxicology, testing drug and ligand receptor interactions (Pramanik 2004; Montet *et al.* 2006; Guryanov *et al.* 2016; Riesenberg *et al.* 2020), and tissue engineering, modeling cells-specific diseases toward the goal of transplantation of healthy cells (Soto-Gutierrez *et al.* 2010).

The surmounting usefulness of cell culture has led to the mass collection and biobanking of cell lines from a variety of species, tissues, and disease states. The National Institute of General Medical Sciences (NIGMS) hosts a repository of over 11,000 cell lines; however, these are predominantly for studies in human medicine. The American Type Culture Collection (ATCC) is a global leader in the preservation, validation, and distribution of a wide range of biological materials, including over 4,000 primary and

continuous cell lines, but only an underwhelming 1% (58) of cell lines are from non-model organisms (ATCC). Although primary culture methods provide an avenue to advance the fields of molecular and evolutionary ecology, these methods have yet to be sufficiently expanded and integrated into the study of natural populations and non-model organism.

Just as the biomedical community is guided by translational research practices, extending basic molecular findings to population-level medicine (Seals 2013; Segeritz and Vallier 2017), evolutionary ecology can exploit this bidirectional research process to extend our molecular knowledge of natural populations to viable population management and conservation practices. Le Pennec & Le Pennec (2007) developed a primary culture model from the commercially relevant scallop *Pecten maximus*, establishing pancreatic acinar cells *in vitro*, for ecotoxicology research focused on the effects of marine pollutants like polycyclic aromatic hydrocarbons (PAHs) from oils and burned carbon (Le Pennec and Le Pennec 2001). Recently, a review and case study of coral cell culture called for more work to be done in characterizing and optimizing *in vitro* culture of coral cells to study bleaching, disease, and toxicity in these complex culture systems. The study established cultures from the coral species, *Pocillopora acuta*, and isolated several host and symbiont cells as well as nematocysts (Roger *et al.* 2021). Moreover, increasingly available cellular resources from natural populations provide a much-needed expansion of the available comparative research models currently used to understand evolutionary and physiological processes and human medicine. There is an established and growing body of research using amphibian and reptile cell models to study regeneration with goals of understanding wound healing in medical research (Lévesque *et al.* 2007; Yokoyama *et*

al. 2018; Franchini 2019). Aging and cancer research has also branched out from the typical biomedical models (Hoekstra *et al.* 2020). Primary fibroblasts have been established from a long-lived and tumor resistant non-model rodent, the naked mole rat (*Heterocephalus glaber*). Evdokimov et al. (2018) used naked mole rat cells and mouse cells in a comparative study of the cellular, transcriptional, and protein modification activity associated with base (BER) and nucleotide (NER) excision repair after exposure to DNA-damaging radiation. The authors found that heightened sensitivity to DNA damage concomitant with higher BER, NER, and poly(ADP-ribose)polymerase (PARP) activity, which correlates with mammalian lifespan (Evdokimov *et al.* 2018).

When considering the benefits of cellular models in the context of animal welfare in research, cell culture embodies the animal research tenet, the "Three R's", - refinement, reduction, and replacement (Fenwick *et al.* 2009). Tissue biopsies for cell culture are relatively small (i.e., ~ 8mm), and can be obtained virtually non-invasively and opportunistically (i.e., obtainment via catch and release sampling, natural autotomy, or postmortem), further refining experimental animal and sample procurement. Primary cell culture can be used to supplement whole organisms in studies, with appropriate research hypotheses, and can reduce the sample sizes and number of studies requiring the use of whole organisms (Nowotny *et al.* 2021).

To successfully integrate primary culture methodology into ecological and evolutionary research, standardized establishment protocols and validated, cryo-banked cell samples are critical. Biopreservation and biobanking of cells from a diversity of

organisms and tissues offers a virtually infinite resources of biomaterials to the scientific community when properly maintained, sub-cultured and published or banked. Expanding cell culture methods used in evolutionary ecology research also benefits related fields of cytotechnology, biotechnology, and omics by providing a sustainable source of cells for the extraction of chromosome-length DNA and high-quality RNA for the preparation of multiple sequencing libraries. Ryder and Onuma (2018) reviewed the applications of cell culture in the characterization and conservation of biodiversity. The authors provide historical and methodological culture information, and they describe the potential for using methods already implemented in model organisms, such karyotyping, genome and phenome sequencing, genetic rescue, and induced pluripotent stem cell (iPSC) generation from *in vitro* culture biomaterials (Ryder and Onuma 2018). The lag in development and publication of these resources for non-model systems is likely due to the extensive time and expertise required for establishing and maintaining primary cells, cell lines, and cell banks or repositories. In addition to trained personnel, the initial cost of establishment and expansion of cell culture resources can easily overwhelm the finances of a laboratory, and funding for the development of these types of resources are often prioritized for medically related research. Towards the goal of promoting primary culture methodology in ecological research, we present methods for the establishment and validation of primary fibroblast cultures from reptiles — multiple lizard species in the *Anolis* genus that is rich with prominent evolutionary and ecological models.

The *Anolis* genus is a well-studied animal system, with well-documented embryonic development (Sanger, Losos, & Gibson-Brown, 2008) and life history knowledge (Lovern *et al.* 2004; Warner and Shine 2008; Warner and Lovern 2014),

evolutionary models for island radiations and niche partitioning (Losos *et al.* 1998; Losos 2011; Angetter *et al.* 2011; Huie *et al.* 2021), and selection for limb size (Sanger *et al.* 2012; Hagey *et al.* 2017). Additionally, there are publicly available genomic and transcriptomic resources available for the brown anole (*Anolis sagrei*) (Geneva et al 2021) and green anole (*Anolis carolinensis*) (Alföldi et al., 2011; Eckalbar et al., 2013). Recently, brown anoles (*A. sagrei*) have been pushed to the forefront as a reptile model for gene editing (Rasys et al., 2019). We detail the establishment primary and early passage cells use with six reptile species, the brown anole (*A. sagrei;* Figure 3.1), green anole (*A. carolinensis*), knight anole (*A. equestris*), large-headed anole (*A. cybotes*), Hispaniolan green anole (*A. chlorocyanus*), and bark anole (*A. distichus*). Details on primary cell validation applies to brown anole cells. These resources and protocols will further endorse anoles as model organisms beyond ecology and evolutionary biology, and ultimately bolster the integration of evolutionary, molecular ecology and biomedical research.

MATERIALS & METHODS

Establishment of Primary and Early Passage Cells from Lizard Tails

Culture protocols are briefly outlined below, but more detailed protocols, recipes, and reagent lists will be published with protocols.io. These methods were developed and optimized through the establishment of ~250 primary cells from 9 lizard species; 123 of

these cells are from the *Anolis* genus (Table S3.1). Here we only present establishment protocols for 5 anole species and validation data from brown anole (*Anolis sagrei)* cells.

Animals and Sample Collection

All primary cells were developed from tail tissues. Anoles, like many lizards, can autotomize and regenerate their tails so establishment can be non-lethal, but all tails used in this study were taken post-mortem. The brown anole (Bahama source population) tails were collected and shipped from The Cox Laboratory (University of Virginia, UVA). 10 minutes after euthanasia via decapitation, tails were severed below the cloaca (~12 - 25mm for female and male, respectively) using a sterile scalpel. The tails were wiped against the scales with an alcohol prep pad or 95% ethanol, before being submerged in 70% ethanol for 5 minutes. Tails were stored in 15 mL conical tubes with Collection media DMEM (4.5g/L glucose; L-glutamine; sodium pyruvate) [VWR], 0.5% Gentamicin (10 mg/mL) [MP Biomedicals], 0.2% Kanamycin (50 mg/mL) [VWR], 0.8% Penicillin/Streptomycin/Amphotericin B 100x mix (10,000 units/mL pen, 10 mg/ml strep, and 25 μg/mL amphoB) [Hyclone] supplemented with 17% 1M HEPES [Hyclone]. Tails from UVA were stored at 4^{o}C up to 2 days before overnight shipment to Auburn University on wet ice. Tails from the green anole, knight anole, and broad-headed anole species were collected from The Warner Laboratory (Auburn University, AU) under IACUC protocol #2021-3875. Tails were collected opportunistically from laboratory colonies during scheduled euthanasia using the protocol describe above except animals were euthanized using MS222, and tails were not shipped. Details on the species,

provenance, and demographics for each individual cells were isolated from are found in Table S3.1.

Explant Methodology

Tails received were stored at 4ºC for 24 -48 hrs. prior to establishment. Cells were established using standard explant culture methods (Polazzi and Alibardi 2011; Freshney 2016) optimized for lizard tissues.

Tails were removed from collection media and submerged in 100% original Listerine [Johnson & Johnson] for 15 – 30 minutes, where the soak time increased with the size of the tail. Tails were transferred to 1X DPBS [VWR] for 2 minutes, prior to removing all scales, bone, cartilage, and adipose tissues with a sterile scalpel. When possible, a portion of the tail was retained and snap frozen for DNA extractions or diced and cryopreserved for future establishment. On a petri dish, the remaining tissue was diced into pieces (~ 10 mm^2) in a small amount of DPBS and collected with sterile forceps into a 1.5mL tube with 700 μL of DPBS. This was lightly vortexed to wash the tissue pieces and reduce the likelihood of transferring contaminants to the establishment plates. This step is most useful for samples that could not be collected aseptically (i.e., in the field). Tissue pieces were transferred to 3 wells of a 6 well plate [VWR] with 500 μL of establishment media (DMEM/F12 1:1(4.5g/L glucose; L-glutamine; sodium pyruvate; HEPES) [Lonza], 16% Fetal Bovine Serum (FBS) [VWR], 3% Chicken Serum (CS) [Equitech-Bio], 1% Non-essential Amino Acids (NEAA) [Hyclone, GE Healthcare], 0.5% Gentamicin, 0.17%, Kanamycin (Optional), 0.8% Penicillin/Streptomycin/AmphotericinB 100X mix) and

diced into pieces > 5mm^2. A small volume of establishment media (< 300 μL) was added to prevent drying out tissue pieces. Establishment plates were incubated at 30ºC with 5% CO_2 in an air or water-jacketed CO_2 incubator [VWR].

Subculturing Methodology

Once a culture reached ~70% confluency, the percentage of surface area covered by cells, media was aspirated from culture dishes and cells were lifted from the dish using 4% trypsin EDTA [Corning] for 3-5 minutes (volume dependent on surface area of dish). Enzymatic activity was neutralized by adding culturing media (DMEM/F12 1:1(4.5g/L glucose; L-glutamine; sodium pyruvate; HEPES), 12% FBS, 3 % CS, 0.5% Gentamicin (Optional), 0.17% Kanamycin (Optional), 0.8% Penicillin/Streptomycin/AmphotericinB mix (Optional) to the dish and using the cell suspension to gently wash any remaining adherent cells from the plate with a 10mL pipette [Thermo Scientific] and electronic pipetman [Argos]. Cell suspension was collected in a conical tube and gently inverted 2 – 3 times before desired volumes of the cell suspension were divided into new culture dishes that already contained the accurate amount of fresh culturing media to bring the total volume with cell suspension to 10mL (appropriate volume for 10 cm dishes).

Cryofreezing and Preservation Methodology

For cryopreservation, media was aspirated from culture dishes and cells were lifted from the dish using 4% trypsin EDTA for 3 -5 minutes. Enzymatic activity was neutralized by

adding media to the dish and using the cell suspension to gently wash any adherent cells from the plate. The cell suspension was collected, and the number of cells was estimated using an EVE automated cell counter [Nanotek]. Cells were pelleted via centrifugation for 5 minutes at 200 xG and culturing media was removed. Freezing media (DMEM/F12 1:1(4.5g/L glucose; L-glutamine; sodium pyruvate), 45% FBS, 5% CS, 10% dimethyl sulfoxide (DMSO) [VWR]) was used to resuspend the cell pellet at a density of 2 million cells per mL. One mL of cells was aliquoted into a 2mL cryotube [Simport] and stored at -80ºC in a 'Mr. Frosty' Cryo 1ºC Freezing Container [Nalgene] that cools at a controlled rate of –1°C/minute for 24 – 48 hrs before long-term storage in vapor phase liquid nitrogen.

Cell Validation & Characterization

Contamination and Mycoplasma Screening

For studies using cell culture systems, validating the health and identity of the cells is imperative to substantiating the validity of the study and findings (Yung 2012). To ensure that all cell cultures were healthy, cultures were visually assessed for cell morphology, growth, and any microbial contaminants using an EVOS XL imaging system [Invitrogen]. *Mycoplasma* species, bacteria that are common culprits of unhealthy cultures, are not visible using light-field microscopy. To screen cultures for the presence of *Mycoplasma* infection, cell supernatant must be lysed, and the resulting solution tested for the DNA of common *Mycoplasma* contaminants. *Mycoplasma* Detection Kit

[Southern Biotechnology], with a detection limit of 2-5 femtograms/100µl supernatant, was used to test for the presence of 19 species from three genera, following the manufacturer's protocol. After cells had been growing successfully in normal growth media, it was supplied with antibiotic-free media and allowed to grow for three to six days and before expended media was collected (1mL). The cell supernatant was centrifuged to lyse any cells, and 100µL were sampled to test for *Mycoplasma* contaminants using PCR-based detection in duplicate reactions. The kit includes a positive control and water was used as a negative control. The reactions were run on a thermal cycler [BioRad T100] using the following program: 95 °C for 5 min, followed by 29 cycles of 95 °C for 30s, 54 °C for 30s, 72 °C for 45s, 60 °C for 30m, and a 12 °C hold. PCR product was visualized using gel electrophoresis on a 0.8% agarose gel in a 50 ml volume with 1.5 µl of GelGreen DNA staining dye [Biotium, USA, Cat. #41005]. Samples were run alongside a 1kb [New England BioLabs, #N0468] and 250 bp [New England BioLabs, #N0557] DNA ladder run at 90V for 45 minutes. The presence of *Mycoplasma* infections is detected by banding between ~448 bp to ~611 bp that matches one of the known amplicon lengths.

Short Tandem Repeat (STR) Profiling for Single Individual Culture Validation

To validate that the established cells were still the identity of the source and not contaminated by another individual we validated the brown anole cells using STR profiling of both the cell pellets and either the tissue or blood taken from the individual at dissection. For each individual, DNA was isolated from the cell pellets and tail tissue or 5

μl of red blood cells taken at euthanization using the PureGeneKit [Qiagen] using the "Cultured Cell Protocol" provided by the manufacture for both the cell pellet and the red blood cells. We used a multiplex PCR to amplify of 5 loci: AAAG-61 (VIC), AAAG-77 (PET), AAAG-91 (FAM), and AAAG-94 (FAM) from Bardeleben et al (2004), and Acar23 (PET) from Wordly et al (2011). The forward primers were ordered with florescent labels as indicated (Integrated DNA Technologies). PCR reactions were conducted in 10 μl volume using forward and reverse primers at a final concentration of 0.5 nM for AAAG-61 and Acar23, 0.075 nM for AAAG-94 and AAAG-77 primers, 0.1 nM for AAAG-91, final concentration of 1X Multiplex PCR master [Qiagen], and 20-50 ng of DNA. The reactions were run on a thermal cycler [BioRad T100] using the following program: 95 °C for 5 min; followed by 29 cycles of 95 °C for 30s, 54 °C for 30s, 72 °C for 45s; a final extension at 60 °C for 30m, and a 12 °C hold. A subsample of the PCR products was run on an 3% agarose gel at 120 voltage for 120 minutes to confirm the amplification of the microsatellites. PCR products were diluted 1:10 and shipped to GeneWiz for fragment analysis on the ABI3730xl. Genotypes were scored using the Microsatellite plugin in the Geneious software v11.1.4 (Kearse et al., 2012), and compared between cell pellets and the blood from the same individuals to verify they match and thereby were not contaminated with cells from other individuals.

Growth curve analysis for growth characterization

To understand the replicative capacity of anole primary fibroblast, two brown anole lines chosen randomly throughout establishment periods were used to calculate

population doubling time. Cells were seeded onto 24 well plates at a density of $1x10^4$ cell/mL and were collected and counted in triplicate for 4 days. Growth curves and calculations for population doubling time were determined using the R package growthcurver. This package fits the data using a standard logistic equation used in population genetics regarding population growth.

RESULTS & DISCUSSION

Here we present optimized protocols for the establishment of reptile primary cells, particularly lizards. Although we only present *Anolis* primary cells here, earlier versions of these methods have also been used to establish primary cell cultures from three additional lizard species (whiptails (*Aspidoscelis* spp.), *Elgaria multicarinata*, *Sceloporus undulatus*; data not included).

Primary cultures are naturally heterogenous populations of cells, with other cell types frequently present early in culture (i.e., keratinocytes, epithelial cells). Medium composition and subculture are both agents of selection for specific cell types. The protocols used for this work were built from published protocols confirmed to be selective for fibroblast cells (Li *et al.* 2016). Microscopy (unstained; brightfield) was used to visually assess cell morphology. Figure 3.1C & D display unstained cells from select sub-confluent and confluent cultures. These fibroblast-like cells have the characteristic bi- and multi-polar morphology seen in sub-confluent fibroblasts, and at confluency, the cells are the expected spindle-like, bipolar shape and organized in whorls.

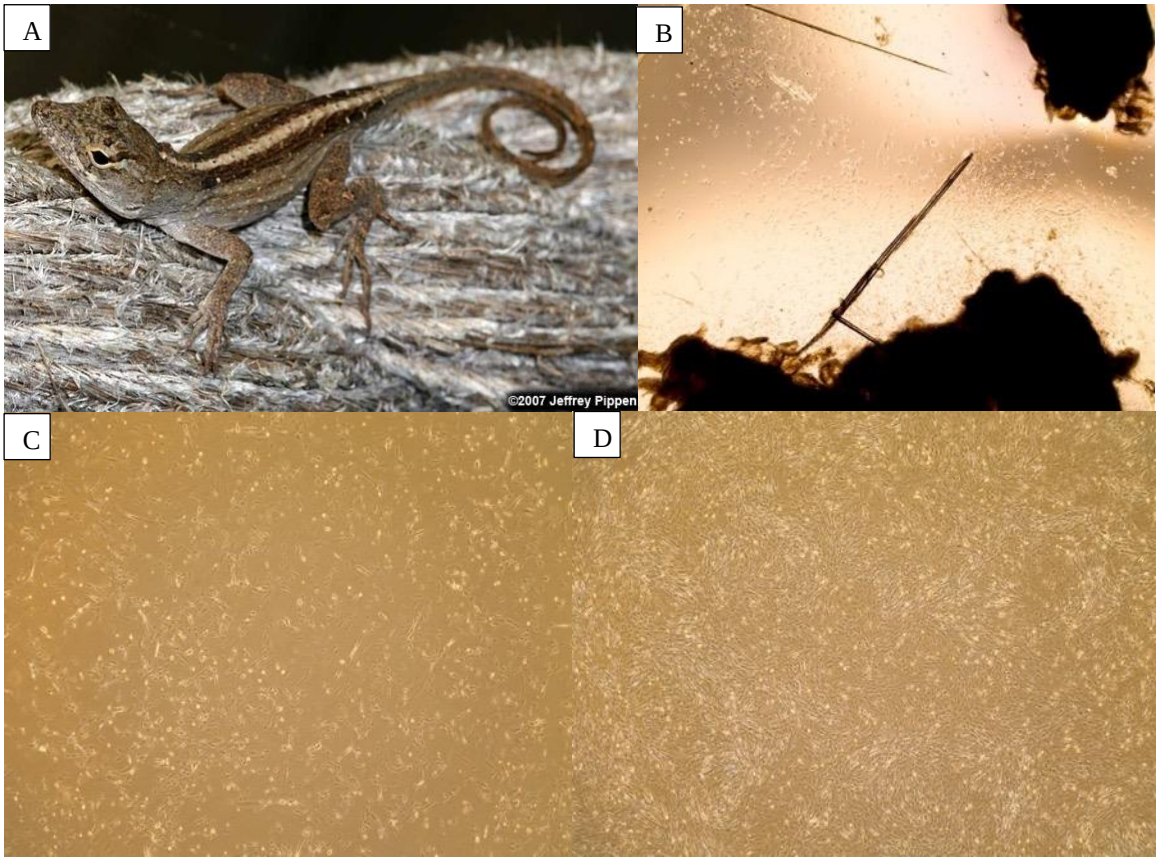

Figure 3.1: Anole cells at different stages of establishment and culture.

Anoles are established evolutionary ecology models. Some species, like the (A) pictured brown anole (*Anolis sagrei*) are being developed to address questions in other fields. Tails are used in (B) explant culture to develop primary and early passage cells. Cells are shown in (C) sub-confluent and (D) confluent states. Cell images were all taken at 4x magnification using an EVOS XL Imaging System by Invitrogen. Photo credit for the anole is to Jeffrey Pippen.

Historically, the misidentification and misrepresentation of cell line identities has resulted in severely conflicting scientific findings, invalidating results generated using cell lines (Horbach and Halffman 2017). The scientific community responded by creating a database of misidentified cell lines and setting a standard for cell culture validation, to ensure that experimental cells are free from microbial, cross-species, and cross-individual contamination (Hughes *et al.* 2007). Along these lines, we tested all cultures for *Mycoplasma* contamination, a common and fatal culture contaminant, with PCR methods. *Mycoplasma* PCR assays were performed prior to cryopreservation. Gel electrophoresis confirmed (1) the expected positive control bands at ~500 bp (*M. orale*) and ~300 bp (internal control), (2) the expected negative control bands at ~300 bp (internal control) and > 75 bp (unused PCR components), and (3) the expected sample bands matching the negative control results, indicating *Mycoplasma*-free cells. Mycoplasma tests should be performed routinely and can be performed using these methods and/or with Hoechst staining.

The identity of all brown anole cultures was validated with genetic data. STR loci previously published and commonly used in brown anole parentage assays were PCR amplified and genotyped for STR profiling. These data were scored based on the genotypes obtained from independent tissue samples that match the origin of the cell sample. Figure 3.2 provides selected scoring data to demonstrate single individual validation. All primary cells successfully profiled demonstrated single individual cultures that matched donor profiles generated from independently sourced DNA, with one exception that was removed from our stocks.

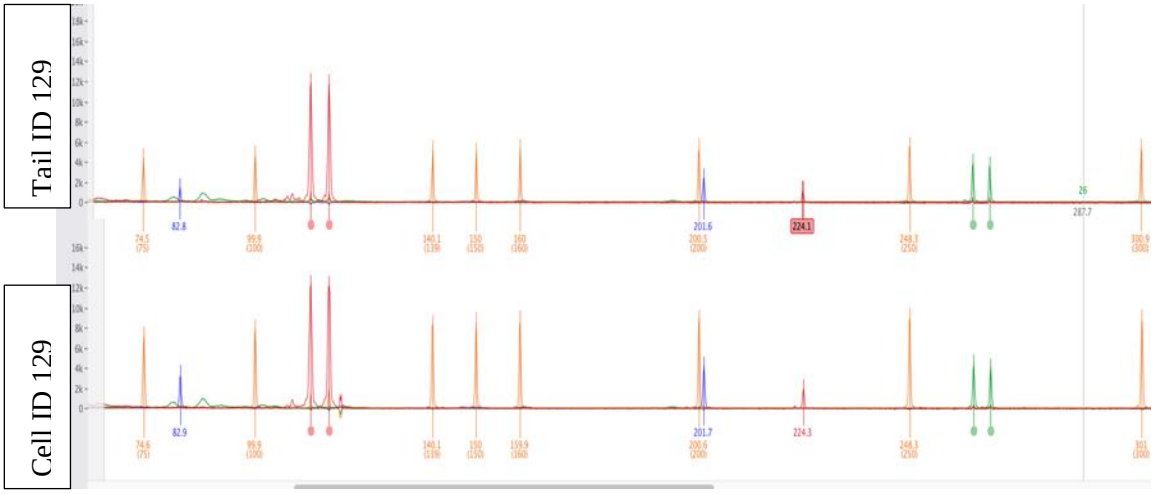

Figure 3.2: Example of Short Tandem Repeat (STR) profiles generated from cell and tail DNA from the same individual match.

Five microsatellite loci (STR) were amplified in a multiplexed PCR using fluorescently labeled primers, and genotyped (blue, red, and green peaks). Orange peaks are internal size ladders. The track in the top panel was generating using DNA extracted from the tail of individual 129 and the bottom panel uses DNA extracted from established cells for the same individual, illustrating matching profiles that indicate the cell culture is representative of the explanted individual.

Growth curve analysis for two primary cells indicate a population doubling time of ~2. 4 days (58 hrs) for brown anole cells. Growth curves in Figure 3.3 demonstrate lag (the time between seeding and the exponential growth phase) times of ~ 2 days. These metrics are important for characterizing and confirming expected growth patterns brown anole cell populations and they demonstrate canonical growth patterns expected for primary cells (Freshney 2016). Finite cells, such as primary and early passage cells typically have longer lag times, relative to continuous cell lines, and population doubling times of 60 – 72 hrs (Freshney 2005). Values estimated from these curves provide a base line expectation to use for detecting changes in growth due to changes in conditions (i.e., responses to different media or other treatments) or cell health (i.e., detect declines in growth due to replicative senescence).

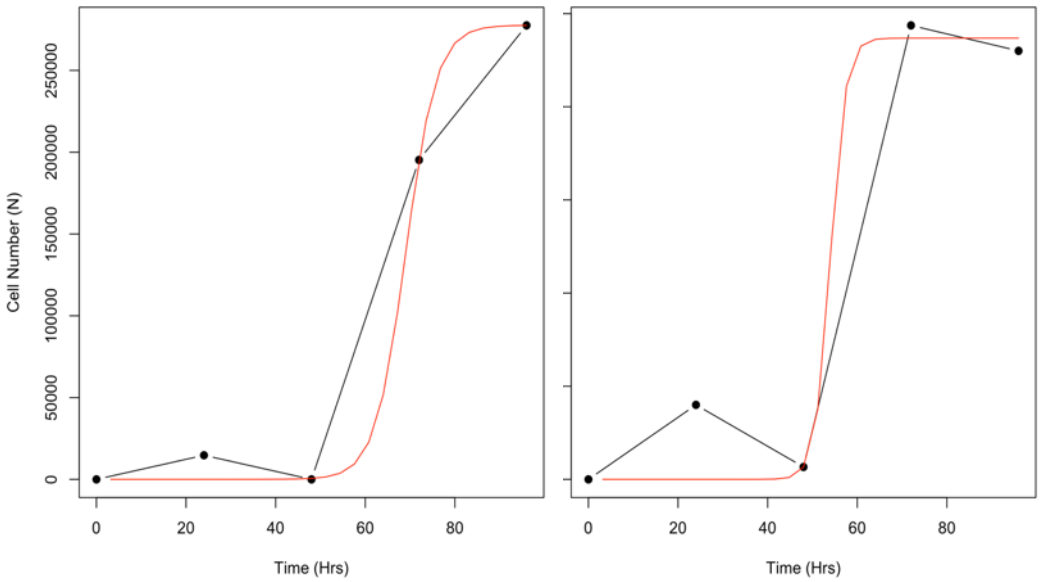

Figure 3.3: Anole primary cells have population doubling times of ~2.4 days.

Cells spend ~ 2 days in the lag phase, prior to exponential growth. Growthcurver fits the data using a common logistics equation. Brown anole cells were plated in 24 well plates at a seeding density of $1x10^4$ cells per well on day 0 and collected for the next four days. Curves are from the 1^{st} and 22^{nd} brown anole cells established.

FUTURE DIRECTIONS & CONCLUSIONS

Fibroblasts are cosmopolitan structural cells important for the formation of the extracellular matrix and wound healing and are relatively easy to select for in culture (Gomes *et al.* 2021). It is important to note the cells isolated here were fibroblast-like but included other cell types at these early culturing stages. *In vitro*, other cell types can also display hallmark fibroblast morphologies, depending on confluency and culturing conditions. Beyond morphological identification, cell type identities should be confirmed using other methods, like staining for cell surface markers. Unfortunately, there are no cell markers exclusively expressed on the surface of fibroblasts across culturing conditions and cell preparations, but probing for vimentin expression and other markers specific to other morphologically similar cell types that can be co-expressed with vimentin (exclusion staining) is common practice for confirming successful isolation of fibroblasts (Goodpaster *et al.* 2008). Although the media preparations presented here are selective for fibroblasts and lack additional factors necessary to support other cell types, these cultures would require molecular cell type characterization and single-clone selection to isolate fibroblasts. It would behoove researchers using these or other methods to identify and or validate cell types in the populations, prior to using primary cells in downstream experiments.

Resources, including methodology, biomaterials, and expertise, are currently limiting factors for the broad inclusion of studies at the cellular level in evolutionary ecology research, making resources like these protocols presented here critical to the field. Culturing costs can also be burdensome for laboratories that don't have existing

equipment, particularly the high initial costs of obtaining CO_2 incubators, biosafety cabinets, and appropriate liquid nitrogen/extremely low temperature storage solutions. Maintenance and consumables costs will vary depending on the scale of the culture system, but a common consumable, fetal bovine serum, cost an average of $400 per 500 mL bottle.

The disparity in available resources for non-model cell culture systems is circular. Specifically, many of the assays used in cellular biology methodologies have been developed using species that are mammalian or tolerant to mammalian culturing conditions, increasing difficulties and efforts to apply them to species outside of these systems. As an example, many protocols that use extended incubation measurements contain enzymes that work optimally at mammalian culturing temperature (37^{o}C). If researchers are using cells that grow at lower temperatures (i.e., reptiles at 28-30^{o}C), they are faced with a decision to culture cells at temperatures that may induce abnormal cell behavior, or to use enzymes at sub-optimal temperatures that may increase assay times or chances of assay failure. As the number and diversity of culture systems developed increases, we can expect more existing resources being optimized or new resources being generated to support non-model cell culture systems.

Researchers new to cellular resources may be curious about the applicability of primary culture methods. Below are two planned uses of the primary cells developed in this work.

1. The role of insulin-like signaling in sex-specific aging using brown anoles

   Brown anole lizards have an average lifespan of 2-4 years (in natural populations) (Reinke *et al.* 2022) and demonstrate consistency in patterns of

gene expression of insulin-like growth factors (IGF1 and IGF2) and sex-specific variation in the context of human aging and longevity (Cox *et al.* 2017; Beatty and Schwartz 2020). Interestingly, these aging and longevity patterns are not consistent in the current vertebrate aging model, the mouse (Soares *et al.* 1985; Stylianopoulou *et al.* 1988). These cells lines will allow us to investigate how the internal sex-specific environments influence patterns of cellular senescence.

2. Evolution of the insulin-like signaling network across anoles

Anoles have a considerable amount of variation in several members of the Insulin and Insulin-like Signaling network (McGaugh *et al.* 2015). Specifically, in regions integral to hormone – receptor binding relationships between the hormones IGF1 and IGF2 and their receptors IGF1R and INSR. These primary cells will allow us to test bioinformatic predictions about how the binding affinity relationships have evolved across these species.

An additional purpose for my development of reptile cell culture methods was to conduct common garden cell culture experiments from mainland and endemic island populations of alligator lizards (*Elgaria*) to address questions on the evolution of body size and for conservation management. The utility of primary cultures in this way exemplifies the importance of non-model cell culture in conservation efforts. Lastly, fibroblast also can be epigenetically reprogrammed into a progenitor cell type called induced pluripotent stem cells (IPSCs). IPSCs are an extremely valuable resource due to their ability to be differentiated into many different cell types, including cardiomyocytes (Li *et al.* 2019),

neurons (Paşca *et al.* 2011), and primordial germ cells (Mitsunaga *et al.* 2019) that can be used to address tissue-specific effects from organisms that have already been sampled.

Incorporating studies at the cellular level can provide foundational, mechanistic knowledge translatable to our understandings of organismal physiology and ecology, and further, of populations and their responses to global change. I intend for this methodology to be applied and improved upon to expand investigations of natural populations to the biological hierarchy of the cell, contributing to our understanding of biodiversity using the smallest unit of life.

Acknowledgements:

We thank Morgan Muell and Dr. Dan Warner from the Warner lab at Auburn University for their knight, large-headed, and green anole tail donations. We thank Dr. Jim Harper for sharing his culturing protocols. This research is supported by NIH-1R15AG064655-01 to TSS and AR.

REFERENCES

Angetter, L.-S., S. Lötters, and D. Rödder, 2011 Climate niche shift in invasive species: the case of the brown anole. Biol. J. Linn. Soc. 104: 943–954.

ATCC ATCC Cell Products.

Beatty, A. E., and T. S. Schwartz, 2020 Gene expression of the IGF hormones and IGF binding proteins across time and tissues in a model reptile. Physiol. Genomics 52: 423–434.

Cox, R. M., C. L. Cox, J. W. McGlothlin, D. C. Card, A. L. Andrew *et al.*, 2017 Hormonally Mediated Increases in Sex-Biased Gene Expression Accompany the Breakdown of Between-Sex Genetic Correlations in a Sexually Dimorphic Lizard. Am. Nat. 189: 315–332.

Evdokimov, A., M. Kutuzov, I. Petruseva, N. Lukjanchikova, E. Kashina *et al.*, 2018 Naked mole rat cells display more efficient excision repair than mouse cells. Aging 10: 1454–1473.

Fenwick, N., G. Griffin, and C. Gauthier, 2009 The welfare of animals used in science: how the "Three Rs" ethic guides improvements. Can. Vet. J. Rev. Veterinaire Can. 50: 523–530.

Fox, A. S., M. Horikawa, and L.-N. L. Ling, 1967 The Use of Drosophila Cell Cultures in Studies of Differentiation. In Vitro 3: 65–84.

Franchini, A., 2019 Adaptive immunity and skin wound healing in amphibian adults. Open Life Sci. 14: 420–426.

Freshney, R. I., 2016 *Culture of Animal Cells: A Manual of Basic Technique and Specialized Applications*. Wiley-Blackwell, Hoboken, New Jersey.

Freshney, R. I., 2005 Quantitation.

Gomes, R. N., F. Manuel, and D. S. Nascimento, 2021 The bright side of fibroblasts: molecular signature and regenerative cues in major organs. Npj Regen. Med. 6: 1–12.

Goodpaster, T., A. Legesse-Miller, M. R. Hameed, S. C. Aisner, J. Randolph-Habecker *et al.*, 2008 An Immunohistochemical Method for Identifying Fibroblasts in

Formalin-fixed, Paraffin-embedded Tissue. J. Histochem. Cytochem. 56: 347–358.

Guryanov, I., S. Fiorucci, and T. Tennikova, 2016 Receptor-ligand interactions: Advanced biomedical applications. Mater. Sci. Eng. C 68: 890–903.

Hagey, T. J., S. Harte, M. Vickers, L. J. Harmon, and L. Schwarzkopf, 2017 There's more than one way to climb a tree: Limb length and microhabitat use in lizards with toe pads. PLOS ONE 12: e0184641.

Hartford Svoboda, K. K., and W. R. Reenstra, 2002 Approaches to studying cellular signaling: A primer for morphologists. Anat. Rec. 269: 123–139.

Hoekstra, L. A., T. S. Schwartz, A. M. Sparkman, D. A. W. Miller, and A. M. Bronikowski, 2020 The untapped potential of reptile biodiversity for understanding how and why animals age. Funct. Ecol. 34: 38–54.

Horbach, S. P. J. M., and W. Halffman, 2017 The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. PLOS ONE 12: e0186281.

Hughes, P., D. Marshall, Y. Reid, H. Parkes, and C. Gelber, 2007 The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? BioTechniques 43: 575–586.

Huie, J. M., I. Prates, R. C. Bell, and K. de Queiroz, 2021 Convergent patterns of adaptive radiation between island and mainland Anolis lizards. Biol. J. Linn. Soc. 134: 85–110.

Keller, G., 2005 Embryonic stem cell differentiation: emergence of a new era in biology and medicine. Genes Dev. 19: 1129–1155.

Keshishian, H., 2004 Ross Harrison's "The outgrowth of the nerve fiber as a mode of protoplasmic movement." J. Exp. Zoolog. A Comp. Exp. Biol. 301A: 201–203.

Landry, J. J. M., P. T. Pyl, T. Rausch, T. Zichner, M. M. Tekkedil *et al.*, 2013 The Genomic and Transcriptomic Landscape of a HeLa Cell Line. G3 GenesGenomesGenetics 3: 1213–1224.

Le Pennec, G., and M. Le Pennec, 2001 Acinar primary cell culture from the digestive gland of Pecten maximus (L.): an original model for ecotoxicological purposes. J. Exp. Mar. Biol. Ecol. 259: 171–187.

Lévesque, M., S. Gatien, K. Finnson, S. Desmeules, É. Villiard *et al.*, 2007 Transforming Growth Factor: β Signaling Is Essential for Limb Regeneration in Axolotls. PLOS ONE 2: e1227.

Li, Y., U. Polak, A. D. Clark, A. D. Bhalla, Y.-Y. Chen *et al.*, 2016 Establishment and Maintenance of Primary Fibroblast Repositories for Rare Diseases—Friedreich's Ataxia Example. Biopreservation Biobanking 14: 324–329.

Li, J., N. Rozwadowska, A. Clark, D. Fil, J. S. Napierala *et al.*, 2019 Excision of the expanded GAA repeats corrects cardiomyopathy phenotypes of iPSC-derived Friedreich's ataxia cardiomyocytes. Stem Cell Res. 40: 101529.

Losos, J. B., 2011 *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles*. Univ of California Press.

Losos, J.B., T.R. Jackman, A. Larson, K. Queiroz, and L. Rodriguez-Schettino, 1998 Contingency and determinism in replicated adaptive radiations of island lizards. Science 279: 2115–2118.

Lovern, M. B., M. M. Holmes, and J. Wade, 2004 The Green Anole (Anolis carolinensis): A Reptilian Model for Laboratory Studies of Reproductive Morphology and Behavior. ILAR J. 45: 54–64.

McGaugh, S. E., A. M. Bronikowski, C.-H. Kuo, D. M. Reding, E. A. Addis *et al.*, 2015 Rapid molecular evolution across amniotes of the IIS/TOR network. Proc. Natl. Acad. Sci. 112: 7055–7060.

Mitsunaga, S., K. Shioda, K. J. Isselbacher, J. H. Hanna, and T. Shioda, 2019 Generation of Human Primordial Germ Cell-like Cells at the Surface of Embryoid Bodies from Primed-pluripotency Induced Pluripotent Stem Cells. J. Vis. Exp. JoVE.

Montet, X., H. Yuan, R. Weissleder, and L. Josephson, 2006 Enzyme-based visualization of receptor–ligand binding in tissues. Lab. Invest. 86: 517–525.

Myhre, J. L., and D. B. Pilgrim, 2010 Cellular differentiation in primary cell cultures from single zebrafish embryos as a model for the study of myogenesis. Zebrafish 7: 255–266.

Nowotny, J. D., M. T. Connelly, and N. Traylor-Knowles, 2021 Novel methods to establish whole-body primary cell cultures for the cnidarians Nematostella vectensis and Pocillopora damicornis. Sci. Rep. 11: 4086.

Paşca, S. P., T. Portmann, I. Voineagu, M. Yazawa, A. Shcheglovitov *et al.*, 2011 Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. Nat. Med. 17: 1657–1662.

Polazzi, E., and L. Alibardi, 2011 Cell culture from lizard skin: A tool for the study of epidermal differentiation. Tissue Cell 43: 350–358.

Pramanik, A., 2004 Ligand-receptor interactions in live cells by fluorescence correlation spectroscopy. Curr. Pharm. Biotechnol. 5: 205–212.

Reinke, B. A., H. Cayuela, F. J. Janzen, J.-F. Lemaître, J.-M. Gaillard *et al.*, 2022 Diverse aging rates in ectothermic tetrapods provide insights for the evolution of aging and longevity. Science 376: 1459–1466.

Riesenberg, C., C. A. Iriarte-Valdez, A. Becker, M. Dienerowitz, A. Heisterkamp *et al.*, 2020 Probing Ligand-Receptor Interaction in Living Cells Using Force Measurements with Optical Tweezers. Front. Bioeng. Biotechnol. 8.

Roger, L. M., H. G. Reich, E. Lawrence, S. Li, W. Vizgaudis *et al.*, 2021 Applying model approaches in non-model systems: A review and case study on coral cell culture. PLOS ONE 16: e0248953.

Ryder, O. A., and M. Onuma, 2018 Viable Cell Culture Banking for Biodiversity Characterization and Conservation. Annu. Rev. Anim. Biosci. 6: 83–98.

Sanger, T. J., L. J. Revell, J. J. Gibson-Brown, and J. B. Losos, 2012 Repeated modification of early limb morphogenesis programmes underlies the convergence of relative limb length in Anolis lizards. Proc. R. Soc. B Biol. Sci. 279: 739–748.

Sastry, S. K., and K. Burridge, 2000 Focal Adhesions: A Nexus for Intracellular Signaling and Cytoskeletal Dynamics. Exp. Cell Res. 261: 25–36.

Saunders, K. B., and P. A. D'Amore, 1992 An in vitro model for cell-cell interactions. Vitro Cell. Dev. Biol. - Anim. 28: 521.

Seals, D. R., 2013 Translational physiology: from molecules to public health. J. Physiol. 591: 3457–3469.

Segeritz, C.-P., and L. Vallier, 2017 Chapter 9 - Cell Culture: Growing Cells as Model Systems In Vitro, pp. 151–172 in *Basic Science Methods for Clinical Researchers*, edited by M. Jalali, F. Y. L. Saldanha, and M. Jalali. Academic Press, Boston.

Soares, M. B., D. N. Ishu, and A. Efstratiadis, 1985 Developmental and tissue-specific expression of a family of transcripts related to rat insulin-like growth factor II mRNA. Nucleic Acids Res. 13: 1119–1134.

Soto-Gutierrez, A., H. Yagi, B. E. Uygun, N. Navarro-Alvarez, K. Uygun *et al.*, 2010 Cell Delivery: From Cell Transplantation to Organ Engineering. Cell Transplant. 19: 655–665.

Stylianopoulou, F., J. Herbert, M. B. Soares, and A. Efstratiadis, 1988 Expression of the insulin-like growth factor II gene in the choroid plexus and the leptomeninges of the adult rat central nervous system. Proc. Natl. Acad. Sci. 85: 141–145.

Warner, D. A., and M. B. Lovern, 2014 The Maternal Environment Affects Offspring Viability via an Indirect Effect of Yolk Investment on Offspring Size. Physiol. Biochem. Zool. 87: 276–287.

Warner, D. A., and R. Shine, 2008 The adaptive significance of temperature-dependent sex determination in a reptile. Nature 451: 566–568.

Yokoyama, H., N. Kudo, M. Todate, Y. Shimada, M. Suzuki *et al.*, 2018 Skin regeneration of amphibians: A novel model for skin regeneration as adults. Dev. Growth Differ. 60: 316–325.

Yung, W. K. A., 2012 The value of cell line validation. Neuro-Oncol. 14: 675.

CONCLUSION

Across this body of work, and my other contributions in Appendix 2, I develop resources for key facets of biological research toward the advancement of knowledge across the fields of ecology, evolutionary biology, and conservation. The collective knowledge from these and adjacent fields contribute to our understanding of the natural world and biodiversity. Biodiversity is extremely important for sustaining our planet and the species that inhabit it. Particularly, we as humans benefit from the services that ecosystems provide us, deemed Nature's Contributions to People (NCP). A subdivision of the UN performed a global assessment and developed 18 categories of NCPs that make it overwhelmingly clear how much we depend on biodiversity, from our food and water sources to our infrastructure and cultural practices (Stange *et al.* 2021). We have gained a plethora of biological knowledge from laboratory and agricultural systems, but their inbred traits or lack of environmental context limits the questions they can answer about natural systems, like how levels of genetic variation in integral or keystone species alters entire ecosystems. Further, traditional model organisms don't fully capture the complexities of life or cannot explain the bounty of Earth's biodiversity. Traditional model organisms are also limited in the biomedical questions they can address. As an example, current vertebrate model organisms completely lack or have limited regenerative capabilities, but several other organisms have full or partial regenerative capabilities throughout their lifespan including amphibians (limbs) and cave fish (heart) (Russell *et al.* 2017; Price *et al.* 2019). Building resources for these taxa can further our understanding of wound healing and improve quality of life for many. Technological

advances are bringing us closer to investigating life in the context of internal (molecular interactions) and external (environmental interactions) complexities, but these efforts need many more resources for generating knowledge, experimental validation, and training investigators in the skills needed for these approaches. My work addresses these gaps by generating resources (methodology, experimental systems, and omics) in multiple non-model organisms for investigations of biodiversity at the genetic level.

More ecological and evolutionary investigations need complementary genetics and omics support. In last 5 years, we have seen increasing numbers of genomes being sequenced for investigating alleles segregating in natural populations of taxa (Unamba *et al.* 2015; Burnett *et al.* 2020). These data have helped us understand several aspects of populations, including how levels of genetic diversity help sustain populations and how they respond to different evolutionary forces. Transcriptomics has bolstered investigations of molecular responses to environmentally relevant conditions, with the option to apply these tools in studies that lack reference genomes suitable for the taxon of interest. New taxa are being discovered using metagenomics, restructuring basal relationships on the Tree of Life (Spang *et al.* 2017). There are still many taxa that are underrepresented in omics data for a variety of reasons, including the sheer volume of currently identified species. I have made genomic contributions, specifically reference genomes for two reptile species (fence lizard, gopher snake) and an aquatic crustacean (water flea; Chapter 1). I have also made contributions in transcriptomics in water fleas (Chapter 2). These resources are important for investigating genetic diversity underlying biodiversity, such as their applications in (1) evolutionary and conservation investigations

of island dwarf gopher snakes and (2) longevity and aquatic ecosystem investigations using water fleas. Biodiversity investigations need expansive resources to support more research that is integrative, holistic, and representative of intricacies of nature. Lowering sequencing costs will continue to make these techniques available to more researchers, extending our investigations to be representative of the diversity of taxa and integrating data from entire communities.

The advances in biodiversity research through omics applications can be overshadowed by gaps in biology education necessary to support the application of these tools and investigating biodiversity. This starts in early life education, with building a sense of stewardship for biodiversity and an understanding of how human life depends on it. I think it is important to foster these relationships in safe, interactive, and fun ways. I do so through outdoor ecology education with the non-profit Fresh Air Family. The mission of the organization is getting families and children outdoors for enjoyable experiences and learning opportunities. During my time at Auburn, I have had the opportunity to provide biodiversity education to children in the Auburn community through summer programs. I also shared this opportunity with Auburn graduate and undergraduate students, training and supporting in teaching of biodiversity to the community. Fostering understanding of and appreciation for biodiversity in the next generation of investigators, policy makers, and entrepreneurs will ensure its perpetuity. Unfortunately, there are also computational and technological hurdles that future and current investigators must face. One of my favorite perspective articles stresses the point that today, all biology is computational biology (Markowetz 2017). The volume of data

and statistical frameworks used to generate hypotheses and make interpretations by coalescing individual insights across biology often require extensive computational skills that are not a part of traditional biology curricula. I have contributed to publicly available educational resources useful in learning R for biologists that is available via GitHub. Chapter 2 addresses two gaps in bioinformatics training resources needed to investigate biodiversity: (1) technical guidance of bioinformatic tools for biologists interpreting gene expression analyses (2) guidance on training investigators (particularly graduate and undergraduate level) to perform RNA-seq analyses for understanding gene expression underlying phenotypes. RNA-seq analyses are informative of the mechanistic underpinnings of phenotypes driving the biological curiosity of several researchers. There are 100's of program options for performing various steps of RNA-seq analyses and making decisions about which combinations of appropriate programs would yield the best results is an important question in the field of bioinformatics and all other fields that apply these techniques. While there have been investigations of different RNA-seq pipelines, my contributions take these analyses a step further into functional pathway enrichment to see how the differences in program choice effect the overall biological interpretations. Understanding the effect of analyses at the level of biological interpretation is the end goal for research questions. Biological research employing omics data has exploded, but there is still a large bottleneck at the point of data management and analysis due to a lack of training or access to that training for most investigators. Demonstrating what parameters truly contribute to differences in final interpretations are important for researchers that are not versed in the many programs available to them. Now, more than ever, it is important that resources like those I have

developed are available to the research community to shrink (or at least help train researchers to jump) the inherent hurdles of working with omics data (Burnett *et al.* 2020). These resources are being prepared to be disseminated via GitHub as a companion to a manuscript submission to a refereed educational resource journal that details how to teach these pipelines as Course-based Research Experiences (CRE). Lowering this barrier, by understanding the variation that contributes to biological interpretations in RNA-seq analyses and by increasing training opportunities, will empower researchers and decrease the time to knowledge dissemination for scientific advancement and inspiration.

More ecological and evolutionary systems need conservation-minded experimental systems. There have been notable successes in establishing cellular resources in non-model organisms, but the skills and facilities necessary to do this work is limited across researchers (Polazzi and Alibardi 2011; Xu *et al.* 2018; Yohe *et al.* 2019; Nowotny *et al.* 2021). There have also been even fewer investigations using these experimental systems, which could be due to a lack of access to cells, methods, or knowledge of their general applicability (Jimenez *et al.* 2013, 2014; Roger *et al.* 2021). One of the listed requirements of being a model organism from the National Institutes of Health (NIH) is having a cell model. My message is not one of making every system fit the NIH standards for model organisms, but instead for researchers in ecology and evolution to take largely whole-organism studies, bioinformatic predictions, and theoretical knowledge into tractable, experimental cellular systems. I generate methodology for developing primary cell culture systems in non-model organisms in

Chapter 3. Primary cell culture systems allow us to test our theoretical and *in silico* predictions with conservation-friendly methods. Cellular studies cannot completely replace whole-organism studies for many questions, but they can reduce the number of organisms needed to confirm findings in more complex settings and guide the experimental design and parameters of those studies. In future work these resources could provide a vital system to investigate species with limitations to traditional whole-organism investigations that require capture and transplantation. As an example, I have contributed to investigations of dwarfism in reptiles on the Channel Islands in California. In addition to the reduction in body size, we documented decreased blood glucose measured at initial capture for island organisms relative to mainland (Sparkman *et al.* 2018). While it would be logical to perform a common garden experiment to understand test for genotype by environment interactions, these are protected organisms that cannot be transplanted. What if we could perform that common garden experiment at the cellular level? We can use cells from programmed resource environments (predictions of high resources on the mainland, and low resources on the islands based on blood glucose setpoints) and culture them in both environments to address our questions without moving a single animal. Molecular ecology research is still in need of other biomaterials like antibodies optimized for non-model molecular investigations and biological assays designed for organisms with different physiologies relative to traditional model organisms. Still, with the protocols I have developed, these types of experiments are now possible and answers to the types of questions like in the previous example are within reach!

My dissertation aims to aid in bridging the gap in resources for fields and research that do not strictly address biomedical inquiries but are instead interested in the larger context of biodiversity. Our continuous investigative efforts have revealed network-like structure in the context of ecological and molecular interactions (Han 2008; Lau *et al.* 2017; Bruder *et al.* 2019; Bechtel 2020). A host of interactions between genes within an organism influence one or more traits. These traits interact with an organism's environment as well as the other individuals and species that share this environment. This means that genetic variation in one population of a species can have significant effects on a population of a different species they share habitat with. Genetic variation across species can interact in both cascade and network-like fashion that shape biodiversity directly and indirectly. Investigating genetic diversity in the context of this high connectivity between life in shared environments is the future direction of research. These systems are what we as humans depend on to sustain our own lives and species (Stange *et al.* 2021). Prior to the expansive and expanding technology and information at the molecular level we have access to now, it made sense to parse out systems to the smallest, controlled questions. Biodiversity research traditionally investigates one focal species, even though the traits and systems present in an environment may be a result of multispecies interactions influencing one another. Building resources like those generated in my work are vital for truly diving into investigation of community genomics, the interaction of genomes through interacting species. Now, we have the technology and background knowledge for many systems we can begin scaffolding information across populations and species to truly investigate mechanisms driving the

interactions, effects, and relationships that shape ecosystems, the benefits we gain from nature, and ultimately biodiversity.

REFERENCES

Anderson, C. B., 2018 Biodiversity monitoring, earth observations and the ecology of scale. Ecol. Lett. 21: 1572–1585.

Ankeny, R. A., and S. Leonelli, 2011 What's so special about model organisms? Stud. Hist. Philos. Sci. Part A 42: 313–323.

Baldridge, D., M. F. Wangler, A. N. Bowman, S. Yamamoto, M. T. Acosta *et al.*, 2021 Model organisms contribute to diagnosis and discovery in the undiagnosed diseases network: current state and a future vision. Orphanet J. Rare Dis. 16: 206.

Barbieri, M., M. Bonafè, C. Franceschi, and G. Paolisso, 2003 Insulin/IGF-I-signaling pathway: an evolutionarily conserved mechanism of longevity from yeast to humans. Am. J. Physiol.-Endocrinol. Metab. 285: E1064–E1071.

Bechtel, W., 2020 Hierarchy and levels: analysing networks to study mechanisms in molecular biology. Philos. Trans. R. Soc. B Biol. Sci. 375: 20190320.

Belmont, J. W., and S. M. Leal, 2005 Complex phenotypes and complex genetics: an introduction to genetic studies of complex traits. Curr. Atheroscler. Rep. 7: 180–187.

Benítez-López, A., L. Santini, J. Gallego-Zamorano, B. Milá, P. Walkden *et al.*, 2021 The island rule explains consistent patterns of body size evolution in terrestrial vertebrates. Nat. Ecol. Evol. 5: 768–786.

Blount, Z. D., 2015 The unexhausted potential of E. coli. eLife 4: e05826.

Bonasio, R., 2015 The expanding epigenetic landscape of non-model organisms (H. H. Hoppeler, Ed.). J. Exp. Biol. 218: 114–122.

Bruder, A., A. Frainer, T. Rota, and R. Primicerio, 2019 The Importance of Ecological Networks in Multiple-Stressor Research and Management. Front. Environ. Sci. 7.

Burnett, K. G., D. S. Durica, D. L. Mykles, J. H. Stillman, and C. Schmidt, 2020 Recommendations for Advancing Genome to Phenome Research in Non-Model Organisms. Integr. Comp. Biol. 60: 397–401.

Card, D. C., R. H. Adams, D. R. Schield, B. W. Perry, A. B. Corbin *et al.*, 2019 Genomic Basis of Convergent Island Phenotypes in Boa Constrictors. Genome Biol. Evol. 11: 3123–3143.

Chatterjee, N., and N. Perrimon, 2021 What fuels the fly: Energy metabolism in Drosophila and its application to the study of obesity and diabetes. Sci. Adv. 7: eabg4336.

Chowański, S., K. Walkowiak-Nowicka, M. Winkiel, P. Marciniak, A. Urbański *et al.*, 2021 Insulin-Like Peptides and Cross-Talk With Other Factors in the Regulation of Insect Metabolism. Front. Physiol. 12.

Christe, P., A. P. Møller, N. Saino, and F. de Lope, 2000 Genetic and environmental components of phenotypic variation in immune response and body size of a colonial bird, Delichon urbica (the house martin). Heredity 85: 75–83.

Denley, A., L. J. Cosgrove, G. W. Booker, J. C. Wallace, and B. E. Forbes, 2005 Molecular interactions of the IGF system. Cytokine Growth Factor Rev. 16: 421–439.

Ellegren, H., 2014 Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29: 51–63.

Fox, R. J., J. M. Donelson, C. Schunter, T. Ravasi, and J. D. Gaitán-Espitia, 2019 Beyond buying time: the role of plasticity in phenotypic adaptation to rapid environmental change. Philos. Trans. R. Soc. B Biol. Sci. 374.

Fujita, S., K. Honda, M. Yamaguchi, S. Fukuzo, T. Saneyasu *et al.*, 2019 Role of Insulin-like Growth Factor-1 in the Central Regulation of Feeding Behavior in Chicks. J. Poult. Sci. 56: 270–276.

Funk, W. C., R. E. Lovich, P. A. Hohenlohe, C. A. Hofman, S. A. Morrison *et al.*, 2016 Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (Urocyon littoralis). Mol. Ecol. 25: 2176–2194.

Gerovasileiou, V., C. Chintiroglou, D. Vafidis, D. Koutsoubas, M. Sini *et al.*, 2015 Census of biodiversity in marine caves of the eastern Mediterranean Sea. Mediterr. Mar. Sci. 245–265.

Glazier, A. M., J. H. Nadeau, and T. J. Aitman, 2002 Finding Genes That Underlie Complex Traits. Science 298: 2345–2349.

Han, J.-D. J., 2008 Understanding biological functions through molecular networks. Cell Res. 18: 224–237.

Herrera-Álvarez, S., E. Karlsson, O. A. Ryder, K. Lindblad-Toh, and A. J. Crawford, 2018 How to make a rodent giant: Genomic basis and tradeoffs of gigantism in the capybara, the world’s largest rodent. bioRxiv 424606.

Hoekstra, H. E., 2006 Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity 97: 222–234.

Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot *et al.*, 2013 The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498–503.

Hsu, F.-C., S. Lindström, J. Sun, F. Wiklund, S.-H. Chen *et al.*, 2008 A multigenic approach to evaluating prostate cancer risk in a systematic replication study. Cancer Genet. Cytogenet. 183: 94–98.

Jimenez, A. G., J. V. Brocklyn, M. Wortman, and J. B. Williams, 2014 Cellular Metabolic Rate Is Influenced by Life-History Traits in Tropical and Temperate Birds. PLOS ONE 9: e87349.

Jimenez, A. G., J. M. Harper, S. A. Queenborough, and J. B. Williams, 2013 Linkages between the life-history evolution of tropical and temperate birds and the resistance of cultured skin fibroblasts to oxidative and non-oxidative chemical injury. J. Exp. Biol. 216: 1373–1380.

Khan, M. J., C. B. Jacometo, D. E. Graugnard, M. N. Corrêa, E. Schmitt *et al.*, 2014 Overfeeding Dairy Cattle during Late-Pregnancy Alters Hepatic PPARα-Regulated Pathways Including Hepatokines: Impact on Metabolism and Peripheral Insulin Sensitivity. Gene Regul. Syst. Biol. 8: GRSB.S14116.

Lau, M. K., S. R. Borrett, B. Baiser, N. J. Gotelli, and A. M. Ellison, 2017 Ecological network metrics: opportunities for synthesis. Ecosphere 8: e01900.

Leonelli, S., and R. A. Ankeny, 2013 What makes a model organism? Endeavour 37: 209–212.

Lomolino, M. V., 2005 Body size evolution in insular vertebrates: generality of the island rule. J. Biogeogr. 32: 1683–1699.

Markowetz, F., 2017 All biology is computational biology. PLOS Biol. 15: e2002050.

Mathieson, I., 2021 The omnigenic model and polygenic prediction of complex traits. Am. J. Hum. Genet. 108: 1558–1563.

Menezes, R., S. Tenreiro, D. Macedo, C. N. Santos, and T. F. Outeiro, 2015 From the baker to the bedside: yeast models of Parkinson's disease. Microb. Cell 2: 262–279.

Metallo, C. M., and M. G. V. Heiden, 2013 Understanding metabolic regulation and its influence on cell physiology. Mol. Cell 49: 388–398.

Muir, P., S. Li, S. Lou, D. Wang, D. J. Spakowicz *et al.*, 2016 The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 17: 53.

Müller, B., and U. Grossniklaus, 2010 Model organisms — A historical perspective. J. Proteomics 73: 2054–2063.

Nowotny, J. D., M. T. Connelly, and N. Traylor-Knowles, 2021 Novel methods to establish whole-body primary cell cultures for the cnidarians Nematostella vectensis and Pocillopora damicornis. Sci. Rep. 11: 4086.

Orr, H. A., 1998 The Population Genetics of Adaptation: The Distribution of Factors Fixed During Adaptive Evolution. Evolution 52: 935–949.

Perry, B. W., A. L. Andrew, A. H. Mostafa Kamal, D. C. Card, D. R. Schield *et al.*, 2019 Multi-species comparisons of snakes identify coordinated signalling networks underlying post-feeding intestinal regeneration. Proc. Biol. Sci. 286: 20190910.

Polazzi, E., and L. Alibardi, 2011 Cell culture from lizard skin: A tool for the study of epidermal differentiation. Tissue Cell 43: 350–358.

Price, E. L., J. M. Vieira, and P. R. Riley, 2019 Model organisms at the heart of regeneration. Dis. Model. Mech. 12: dmm040691.

Radwan, J., and W. Babik, 2012 The genomics of adaptation. Proc. R. Soc. B Biol. Sci. 279: 5024–5028.

Reuter, J. A., D. V. Spacek, and M. P. Snyder, 2015 High-Throughput Sequencing Technologies. Mol. Cell 58: 586–597.

Roger, L. M., H. G. Reich, E. Lawrence, S. Li, W. Vizgaudis *et al.*, 2021 Applying model approaches in non-model systems: A review and case study on coral cell culture. PLOS ONE 16: e0248953.

Russell, J. J., J. A. Theriot, P. Sood, W. F. Marshall, L. F. Landweber *et al.*, 2017 Non-model model organisms. BMC Biol. 15: 55.

Schlichting, C. D., and H. Smith, 2002 Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. Evol. Ecol. 16: 189–211.

Schwander, T., R. Libbrecht, and L. Keller, 2014 Supergenes and Complex Phenotypes. Curr. Biol. 24: R288–R294.

Seifirad, S., and V. Haghpanah, 2019 Inappropriate modeling of chronic and complex disorders: How to reconsider the approach in the context of predictive, preventive and personalized medicine, and translational medicine. EPMA J. 10: 195–209.

Soon, W., M. Hariharan, and M. Synder, 2013 High-throughput sequencing for biology and medicine. Mol. Syst. Biol. 9: 640.

Spang, A., E. F. Caceres, and T. J. G. Ettema, 2017 Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science 357: eaaf3883.

Sparkman, A. M., A. D. Clark, L. J. Brummett, K. R. Chism, L. L. Combrink *et al.*, 2018 Convergence in reduced body size, head size, and blood glucose in three island reptiles. Ecol. Evol. 8: 6169–6182.

Stange, M., R. D. H. Barrett, and A. P. Hendry, 2021 The importance of genomic variation for biodiversity, ecosystems and people. Nat. Rev. Genet. 22: 89–105.

Stockdale, W. T., M. E. Lemieux, A. C. Killen, J. Zhao, Z. Hu *et al.*, 2018 Heart Regeneration in the Mexican Cavefish. Cell Rep. 25: 1997-2007.e7.

Székely, T., 2019 Why study plovers? The significance of non-model organisms in avian ecology, behaviour and evolution. J. Ornithol. 160: 923–933.

Unamba, C. I. N., A. Nag, and R. K. Sharma, 2015 Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. Front. Plant Sci. 6:.

Veldman, M. B., and S. Lin, 2008 Zebrafish as a Developmental Model Organism for Pediatric Research. Pediatr. Res. 64: 470–476.

Voultsiadou, E., V. Gerovasileiou, L. Vandepitte, K. Ganias, and C. Arvanitidis, 2017 Aristotle's scientific contributions to the classification, nomenclature and distribution of marine organisms. Mediterr. Mar. Sci. 18: 468–478.

Vucetich, J. A., E. A. Macdonald, D. Burnham, J. T. Bruskotter, D. D. P. Johnson *et al.*, 2021 Finding Purpose in the Conservation of Biodiversity by the Commingling of Science and Ethics. Animals 11: 837.

Xu, H., X. Zhu, W. Li, Z. Tang, Y. Zhao *et al.*, 2018 Isolation and in vitro culture of ovarian stem cells in Chinese soft-shell turtle (Pelodiscus sinensis). J. Cell. Biochem. 119: 7667–7677.

Yohe, L. R., P. Devanna, K. T. J. Davies, J. H. T. Potter, S. J. Rossiter *et al.*, 2019 Tissue Collection of Bats for -Omics Analyses and Primary Cell Culture. JoVE J. Vis. Exp. e59505.

APPENDIX 1: Supplementary Material

Chapter 1

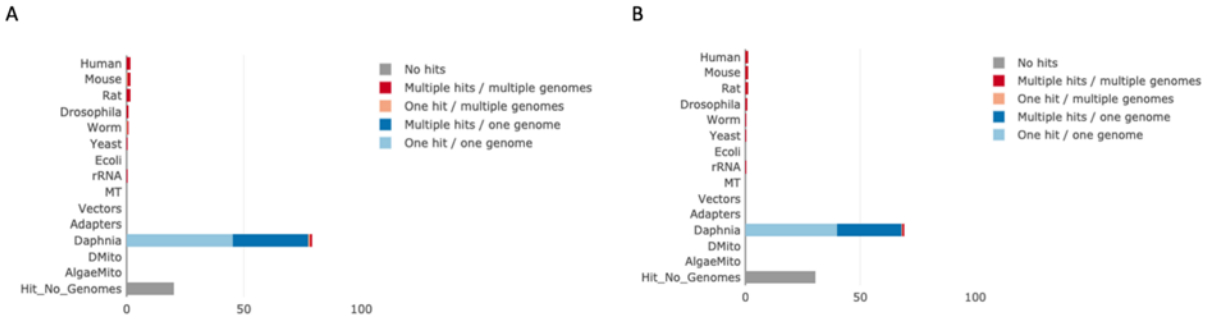

Figure S1.1: FastQC Screen Analysis Indicates the Expected Composition of Reads Based on Screened Genomes. Read subsets from each *D. pulicaria* library were mapped against several common and selected sequencing contaminants using bowtie2. Plots indicate that majority of the reads are mapping uniquely to the bait genome PA42 ("*Daphnia*"), as expected. For other genomes in the search library, the reads did not map uniquely and likely represent low-complexity regions. There is a significant proportion of the read subset that does not map to any represented sequences represented with grey bars.

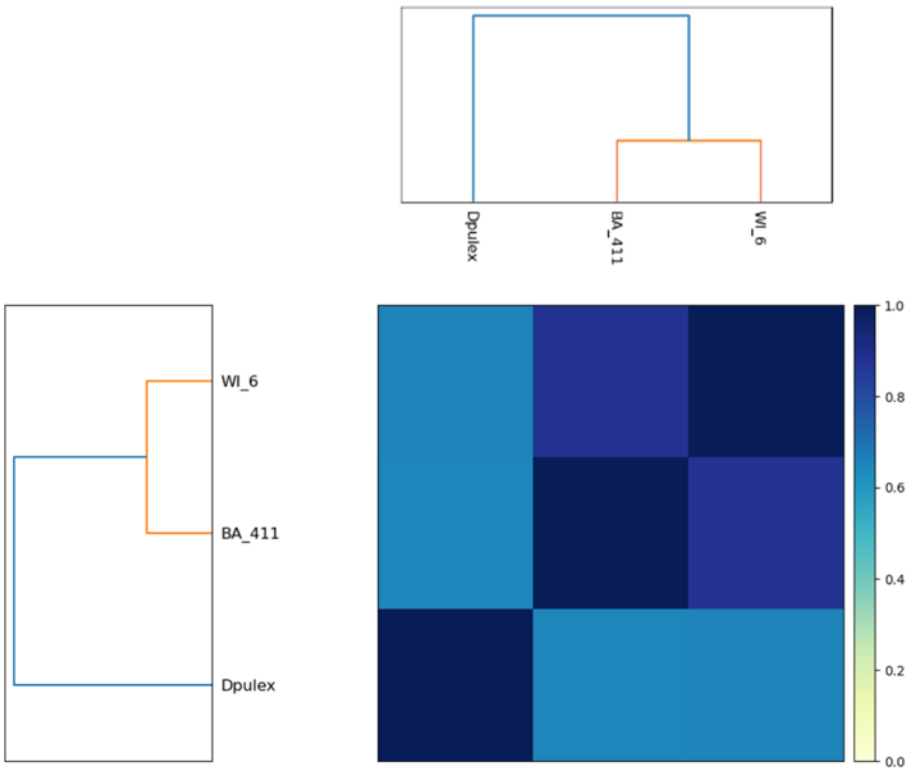

Figure S1.2: Sourmash Distance Estimates Indicate Higher Similarity Between *D. pulicaria* Strains (BA411 & WI6), Relative to the *D. pulex* (PA42) Reference. Dendrograms on the top and left recapitulate the relationship between samples in the distance analysis that is also visualized by the matrix. The color gradient indicates the sourmash distance estimate, where darker colors indicate high similarity between samples and lighter colors indicate more divergent samples. The *D. pulicaria* assemblies only vary by 0.1 when compared to each other.

Chapter 2

Table S2.1 Contrast between pipelines from emmeans for Model 1. Pairwise contrasts between each pipeline. Estimates are differences between predicted (estimated marginal) means for pipeline comparisons. Contrasts have been filtered for significant values less than or equal to 0.05 to reduce table sizes. SE – standard error; see Table 2.1 for program abbreviations.

Model 1 (Soft Filtered):

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_St DESeq2 - Hi_Ht DESeq2	-75.000	11.328	10	-125.777	-24.222	-6.620	0.003
K_K2 DESeq2 - Hi_St DESeq2	54	11.328	10	3.222	104.777	4.766	0.034
Sl_Sl2 DESeq2 - Hi_St DESeq2	51.666	11.328	10	0.889	102.443	4.560	0.045
Sr_Ht DESeq2 - Hi_St DESeq2	80.333	11.328	10	29.556	131.110	7.091	0.001
Sr_St DESeq2 - Sr_Ht DESeq2	-53.333	11.328	10	-104.110	-2.556	-4.707	0.036
Hi_Ht edgeR - Hi_Ht DESeq2	55.833	8.010	10	19.928	91.738	6.970	0.002
Hi_Ht edgeR - Hi_St DESeq2	130.833	13.874	10	68.644	193.022	9.429	0.000
Hi_Ht edgeR - K_K2 DESeq2	76.833	13.874	10	14.644	139.022	5.537	0.012
Hi_Ht edgeR - Sl_Sl2 DESeq2	79.166	13.874	10	16.977	141.355	5.705	0.010

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_Ht edgeR - Sr_St DESeq2	103.833	13.874	10	41.644	166.022	7.483	0.001
Hi_St edgeR - Hi_St DESeq2	55.833	8.010	10	19.928	91.738	6.970	0.002
Hi_St edgeR - Hi_Ht edgeR	-75.000	11.328	10	-125.777	-24.222	-6.620	0.003
K_K2 edgeR - Hi_St DESeq2	109.833	13.874	10	47.644	172.022	7.916	0.000
K_K2 edgeR - K_K2 DESeq2	55.833	8.010	10	19.928	91.738	6.970	0.002
K_K2 edgeR - Sr_St DESeq2	82.833	13.874	10	20.644	145.022	5.970	0.007
K_K2 edgeR - Hi_St edgeR	54	11.328	10	3.222	104.777	4.766	0.034
Sl_Sl2 edgeR - Hi_St DESeq2	107.5	13.874	10	45.310	169.689	7.748	0.000
Sl_Sl2 edgeR - Sl_Sl2 DESeq2	55.833	8.010	10	19.928	91.738	6.970	0.002
Sl_Sl2 edgeR - Sr_St DESeq2	80.499	13.874	10	18.310	142.689	5.802	0.008
Sl_Sl2 edgeR - Hi_St edgeR	51.666	11.328	10	0.889	102.443	4.560	0.045
Sr_Ht edgeR - Hi_St DESeq2	136.166	13.874	10	73.977	198.355	9.814	0.000
Sr_Ht edgeR - K_K2 DESeq2	82.166	13.874	10	19.977	144.355	5.922	0.007
Sr_Ht edgeR - Sl_Sl2 DESeq2	84.499	13.874	10	22.310	146.689	6.090	0.006
Sr_Ht edgeR - Sr_Ht DESeq2	55.833	8.010	10	19.928	91.738	6.970	0.002
Sr_Ht edgeR - Sr_St DESeq2	109.166	13.874	10	46.977	171.355	7.868	0.000
Sr_Ht edgeR - Hi_St edgeR	80.333	11.328	10	29.556	131.110	7.091	0.001

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_St edgeR - Hi_St DESeq2	82.833	13.874	10	20.644	145.022	5.970	0.007
Sr_St edgeR - Sr_St DESeq2	55.833	8.010	10	19.928	91.738	6.970	0.002
Sr_St edgeR - Sr_Ht edgeR	-53.333	11.328	10	-104.110	-2.556	-4.707	0.036
Hi_Ht LimmaVoom - Hi_Ht DESeq2	156.166	8.010	10	120.261	192.071	19.495	2.249e-07
Hi_Ht LimmaVoom - Hi_St DESeq2	231.166	13.874	10	168.977	293.355	16.661	8.800e-07
Hi_Ht LimmaVoom - K_K2 DESeq2	177.166	13.874	10	114.977	239.355	12.769	1.061e-05
Hi_Ht LimmaVoom - Sl_Sl2 DESeq2	179.5	13.874	10	117.310	241.689	12.937	9.361e-06
Hi_Ht LimmaVoom - Sr_Ht DESeq2	150.833	13.874	10	88.644	213.022	10.871	4.730e-05
Hi_Ht LimmaVoom - Sr_St DESeq2	204.166	13.874	10	141.977	266.355	14.715	2.684e-06
Hi_Ht LimmaVoom - Hi_Ht edgeR	100.333	8.010	10	64.428	136.238	12.525	1.276e-05
Hi_Ht LimmaVoom - Hi_St edgeR	175.333	13.874	10	113.144	237.522	12.637	1.172e-05
Hi_Ht LimmaVoom - K_K2 edgeR	121.333	13.874	10	59.144	183.522	8.745	0.000
Hi_Ht LimmaVoom - Sl_Sl2 edgeR	123.666	13.874	10	61.477	185.855	8.913	0.000
Hi_Ht LimmaVoom - Sr_Ht edgeR	95.000	13.874	10	32.810	157.189	6.847	0.002
Hi_Ht LimmaVoom - Sr_St edgeR	148.333	13.874	10	86.144	210.522	10.691	5.502e-05
Hi_St LimmaVoom - Hi_Ht DESeq2	81.166	13.874	10	18.977	143.355	5.850	0.008
Hi_St LimmaVoom - Hi_St DESeq2	156.166	8.010	10	120.261	192.071	19.495	2.249e-07

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_St LimmaVoom - K_K2 DESeq2	102.166	13.874	10	39.977	164.355	7.363	0.001
Hi_St LimmaVoom - Sl_Sl2 DESeq2	104.5	13.874	10	42.310	166.689	7.531	0.001
Hi_St LimmaVoom - Sr_Ht DESeq2	75.833	13.874	10	13.644	138.022	5.465	0.013
Hi_St LimmaVoom - Sr_St DESeq2	129.166	13.874	10	66.977	191.355	9.309	0.000
Hi_St LimmaVoom - Hi_St edgeR	100.333	8.010	10	64.428	136.238	12.525	1.276e-05
Hi_St LimmaVoom - Sr_St edgeR	73.333	13.874	10	11.144	135.522	5.285	0.017
Hi_St LimmaVoom - Hi_Ht LimmaVoom	-75.000	11.328	10	-125.777	-24.222	-6.620	0.003
K_K2 LimmaVoom - Hi_Ht DESeq2	135.166	13.874	10	72.977	197.355	9.742	0.000
K_K2 LimmaVoom - Hi_St DESeq2	210.166	13.874	10	147.977	272.355	15.147	2.043e-06
K_K2 LimmaVoom - K_K2 DESeq2	156.166	8.010	10	120.261	192.071	19.495	2.249e-07
K_K2 LimmaVoom - Sl_Sl2 DESeq2	158.5	13.874	10	96.310	220.689	11.423	3.011e-05
K_K2 LimmaVoom - Sr_Ht DESeq2	129.833	13.874	10	67.644	192.022	9.357	0.000
K_K2 LimmaVoom - Sr_St DESeq2	183.166	13.874	10	120.977	245.355	13.201	7.699e-06
K_K2 LimmaVoom - Hi_Ht edgeR	79.333	13.874	10	17.144	141.522	5.717	0.009
K_K2 LimmaVoom - Hi_St edgeR	154.333	13.874	10	92.144	216.522	11.123	3.840e-05
K_K2 LimmaVoom - K_K2 edgeR	100.333	8.010	10	64.428	136.238	12.525	1.276e-05
K_K2 LimmaVoom - Sl_Sl2 edgeR	102.666	13.874	10	40.477	164.855	7.399	0.001

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
K_K2 LimmaVoom - Sr_Ht edgeR	74	13.874	10	11.810	136.189	5.333	0.016
K_K2 LimmaVoom - Sr_St edgeR	127.333	13.874	10	65.144	189.522	9.177	0.000
K_K2 LimmaVoom - Hi_St LimmaVoom	54	11.328	10	3.222	104.777	4.766	0.034
Sl_Sl2 LimmaVoom - Hi_Ht DESeq2	132.833	13.874	10	70.644	195.022	9.573	0.000
Sl_Sl2 LimmaVoom - Hi_St DESeq2	207.833	13.874	10	145.644	270.022	14.979	2.268e-06
Sl_Sl2 LimmaVoom - K_K2 DESeq2	153.833	13.874	10	91.644	216.022	11.087	3.955e-05
Sl_Sl2 LimmaVoom - Sl_Sl2 DESeq2	156.166	8.010	10	120.261	192.071	19.495	2.249e-07
Sl_Sl2 LimmaVoom - Sr_Ht DESeq2	127.5	13.874	10	65.310	189.689	9.189	0.000
Sl_Sl2 LimmaVoom - Sr_St DESeq2	180.833	13.874	10	118.644	243.022	13.033	8.716e-06
Sl_Sl2 LimmaVoom - Hi_Ht edgeR	76.999	13.874	10	14.810	139.189	5.549	0.012
Sl_Sl2 LimmaVoom - Hi_St edgeR	152	13.874	10	89.810	214.189	10.955	4.411e-05
Sl_Sl2 LimmaVoom - K_K2 edgeR	98	13.874	10	35.810	160.189	7.063	0.001
Sl_Sl2 LimmaVoom - Sl_Sl2 edgeR	100.333	8.010	10	64.428	136.238	12.525	1.276e-05
Sl_Sl2 LimmaVoom - Sr_Ht edgeR	71.666	13.874	10	9.477	133.855	5.165	0.020
Sl_Sl2 LimmaVoom - Sr_St edgeR	125	13.874	10	62.810	187.189	9.009	0.000
Sl_Sl2 LimmaVoom - Hi_St LimmaVoom	51.666	11.328	10	0.889	102.443	4.560	0.045
Sr_Ht LimmaVoom - Hi_Ht DESeq2	161.5	13.874	10	99.310	223.689	11.640	2.534e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_Ht LimmaVoom - Hi_St DESeq2	236.5	13.874	10	174.310	298.689	17.045	7.267e-07
Sr_Ht LimmaVoom - K_K2 DESeq2	182.5	13.874	10	120.310	244.689	13.153	7.976e-06
Sr_Ht LimmaVoom - Sl_Sl2 DESeq2	184.833	13.874	10	122.644	247.022	13.321	7.051e-06
Sr_Ht LimmaVoom - Sr_Ht DESeq2	156.166	8.010	10	120.261	192.071	19.495	2.249e-07
Sr_Ht LimmaVoom - Sr_St DESeq2	209.5	13.874	10	147.310	271.689	15.099	2.104e-06
Sr_Ht LimmaVoom - Hi_Ht edgeR	105.666	13.874	10	43.477	167.855	7.615	0.001
Sr_Ht LimmaVoom - Hi_St edgeR	180.666	13.874	10	118.477	242.855	13.021	8.794e-06
Sr_Ht LimmaVoom - K_K2 edgeR	126.666	13.874	10	64.477	188.855	9.129	0.000
Sr_Ht LimmaVoom - Sl_Sl2 edgeR	129	13.874	10	66.810	191.189	9.297	0.000
Sr_Ht LimmaVoom - Sr_Ht edgeR	100.333	8.010	10	64.428	136.238	12.525	1.276e-05
Sr_Ht LimmaVoom - Sr_St edgeR	153.666	13.874	10	91.477	215.855	11.075	3.995e-05
Sr_Ht LimmaVoom - Hi_St LimmaVoom	80.333	11.328	10	29.556	131.110	7.091	0.001
Sr_St LimmaVoom - Hi_Ht DESeq2	108.166	13.874	10	45.977	170.355	7.796	0.000
Sr_St LimmaVoom - Hi_St DESeq2	183.166	13.874	10	120.977	245.355	13.201	7.699e-06
Sr_St LimmaVoom - K_K2 DESeq2	129.166	13.874	10	66.977	191.355	9.309	0.000
Sr_St LimmaVoom - Sl_Sl2 DESeq2	131.5	13.874	10	69.310	193.689	9.477	0.000
Sr_St LimmaVoom - Sr_Ht DESeq2	102.833	13.874	10	40.644	165.022	7.411	0.001

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_St LimmaVoom - Sr_St DESeq2	156.166	8.010	10	120.261	192.071	19.495	2.249e-07
Sr_St LimmaVoom - Hi_St edgeR	127.333	13.874	10	65.144	189.522	9.177	0.000
Sr_St LimmaVoom - K_K2 edgeR	73.333	13.874	10	11.144	135.522	5.285	0.017
Sr_St LimmaVoom - Sl_Sl2 edgeR	75.666	13.874	10	13.477	137.855	5.453	0.013
Sr_St LimmaVoom - Sr_St edgeR	100.333	8.010	10	64.428	136.238	12.525	1.276e-05
Sr_St LimmaVoom - Sr_Ht LimmaVoom	-53.333	11.328	10	-104.110	-2.556	-4.707	0.036

Model 1 (Hard Filtered):

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_St DESeq2 - Hi_Ht DESeq2	-77.666	15.653	10	-147.831	-7.501	-4.961	0.026
Sr_St DESeq2 - Hi_Ht DESeq2	-88.000	15.653	10	-158.165	-17.834	-5.621	0.011
Hi_Ht edgeR - Hi_St DESeq2	115.666	19.172	10	29.732	201.601	6.033	0.006
Hi_Ht edgeR - Sr_St DESeq2	126	19.172	10	40.065	211.934	6.572	0.003
Hi_St edgeR - Hi_Ht edgeR	-77.666	15.653	10	-147.831	-7.501	-4.961	0.026
K_K2 edgeR - Sr_St DESeq2	88.333	19.172	10	2.398	174.267	4.607	0.042

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sl_Sl2 edgeR - Sr_St DESeq2	89.666	19.172	10	3.732	175.601	4.676	0.038
Sr_Ht edgeR - Hi_St DESeq2	97.333	19.172	10	11.398	183.267	5.076	0.022
Sr_Ht edgeR - Sr_St DESeq2	107.666	19.172	10	21.732	193.601	5.615	0.011
Sr_St edgeR - Hi_Ht edgeR	-88.000	15.653	10	-158.165	-17.834	-5.621	0.011
Hi_Ht LimmaVoom - Hi_Ht DESeq2	134.5	11.068	10	84.885	184.114	12.151	1.699e-05
Hi_Ht LimmaVoom - Hi_St DESeq2	212.166	19.172	10	126.232	298.101	11.066	4.024e-05
Hi_Ht LimmaVoom - K_K2 DESeq2	172.166	19.172	10	86.232	258.101	8.980	0.000
Hi_Ht LimmaVoom - Sl_Sl2 DESeq2	170.833	19.172	10	84.898	256.767	8.910	0.000
Hi_Ht LimmaVoom - Sr_Ht DESeq2	152.833	19.172	10	66.898	238.767	7.971	0.000
Hi_Ht LimmaVoom - Sr_St DESeq2	222.5	19.172	10	136.565	308.434	11.605	2.604e-05
Hi_Ht LimmaVoom - Hi_Ht edgeR	96.5	11.068	10	46.885	146.114	8.718	0.000
Hi_Ht LimmaVoom - Hi_St edgeR	174.166	19.172	10	88.232	260.101	9.084	0.000
Hi_Ht LimmaVoom - K_K2 edgeR	134.166	19.172	10	48.232	220.101	6.998	0.002
Hi_Ht LimmaVoom - Sl_Sl2 edgeR	132.833	19.172	10	46.898	218.767	6.928	0.002
Hi_Ht LimmaVoom - Sr_Ht edgeR	114.833	19.172	10	28.898	200.767	5.989	0.007
Hi_Ht LimmaVoom - Sr_St edgeR	184.500	19.172	10	98.565	270.434	9.623	0.000
Hi_St LimmaVoom - Hi_St DESeq2	134.5	11.068	10	84.885	184.114	12.151	1.699e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_St LimmaVoom - K_K2 DESeq2	94.499	19.172	10	8.565	180.434	4.929	0.027
Hi_St LimmaVoom - Sl_Sl2 DESeq2	93.166	19.172	10	7.232	179.101	4.859	0.030
Hi_St LimmaVoom - Sr_St DESeq2	144.833	19.172	10	58.898	230.767	7.554	0.001
Hi_St LimmaVoom - Hi_St edgeR	96.5	11.068	10	46.885	146.114	8.718	0.000
Hi_St LimmaVoom - Sr_St edgeR	106.833	19.172	10	20.898	192.767	5.572	0.011
Hi_St LimmaVoom - Hi_Ht LimmaVoom	-77.666	15.653	10	-147.831	-7.501	-4.961	0.026
K_K2 LimmaVoom - Hi_Ht DESeq2	96.833	19.172	10	10.898	182.767	5.050	0.023
K_K2 LimmaVoom - Hi_St DESeq2	174.5	19.172	10	88.565	260.434	9.101	0.000
K_K2 LimmaVoom - K_K2 DESeq2	134.5	11.068	10	84.885	184.114	12.151	1.699e-05
K_K2 LimmaVoom - Sl_Sl2 DESeq2	133.166	19.172	10	47.232	219.101	6.945	0.002
K_K2 LimmaVoom - Sr_Ht DESeq2	115.166	19.172	10	29.232	201.101	6.007	0.006
K_K2 LimmaVoom - Sr_St DESeq2	184.833	19.172	10	98.898	270.767	9.640	0.000
K_K2 LimmaVoom - Hi_St edgeR	136.5	19.172	10	50.565	222.434	7.119	0.001
K_K2 LimmaVoom - K_K2 edgeR	96.5	11.068	10	46.885	146.114	8.718	0.000
K_K2 LimmaVoom - Sl_Sl2 edgeR	95.166	19.172	10	9.232	181.101	4.963	0.026
K_K2 LimmaVoom - Sr_St edgeR	146.833	19.172	10	60.898	232.767	7.658	0.001
Sl_Sl2 LimmaVoom - Hi_Ht DESeq2	98.166	19.172	10	12.232	184.101	5.120	0.021

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sl_Sl2 LimmaVoom - Hi_St DESeq2	175.833	19.172	10	89.898	261.767	9.171	0.000
Sl_Sl2 LimmaVoom - K_K2 DESeq2	135.833	19.172	10	49.898	221.767	7.084	0.001
Sl_Sl2 LimmaVoom - Sl_Sl2 DESeq2	134.5	11.068	10	84.885	184.114	12.151	1.699e-05
Sl_Sl2 LimmaVoom - Sr_Ht DESeq2	116.5	19.172	10	30.565	202.434	6.076	0.006
Sl_Sl2 LimmaVoom - Sr_St DESeq2	186.166	19.172	10	100.232	272.101	9.710	0.000
Sl_Sl2 LimmaVoom - Hi_St edgeR	137.833	19.172	10	51.898	223.767	7.189	0.001
Sl_Sl2 LimmaVoom - K_K2 edgeR	97.833	19.172	10	11.898	183.767	5.102	0.021
Sl_Sl2 LimmaVoom - Sl_Sl2 edgeR	96.5	11.068	10	46.885	146.114	8.718	0.000
Sl_Sl2 LimmaVoom - Sr_St edgeR	148.166	19.172	10	62.232	234.101	7.728	0.000
Sr_Ht LimmaVoom - Hi_Ht DESeq2	116.166	19.172	10	30.232	202.101	6.059	0.006
Sr_Ht LimmaVoom - Hi_St DESeq2	193.833	19.172	10	107.898	279.767	10.110	9.090e-05
Sr_Ht LimmaVoom - K_K2 DESeq2	153.833	19.172	10	67.898	239.767	8.023	0.000
Sr_Ht LimmaVoom - Sl_Sl2 DESeq2	152.5	19.172	10	66.565	238.434	7.954	0.000
Sr_Ht LimmaVoom - Sr_Ht DESeq2	134.5	11.068	10	84.885	184.114	12.151	1.699e-05
Sr_Ht LimmaVoom - Sr_St DESeq2	204.166	19.172	10	118.232	290.101	10.649	5.700e-05
Sr_Ht LimmaVoom - Hi_St edgeR	155.833	19.172	10	69.898	241.767	8.128	0.000
Sr_Ht LimmaVoom - K_K2 edgeR	115.833	19.172	10	29.898	201.767	6.041	0.006

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_Ht LimmaVoom - Sl_Sl2 edgeR	114.5	19.172	10	28.565	200.434	5.972	0.007
Sr_Ht LimmaVoom - Sr_Ht edgeR	96.5	11.068	10	46.885	146.114	8.718	0.000
Sr_Ht LimmaVoom - Sr_St edgeR	166.166	19.172	10	80.232	252.101	8.667	0.000
Sr_St LimmaVoom - Hi_St DESeq2	124.166	19.172	10	38.232	210.101	6.476	0.003
Sr_St LimmaVoom - Sr_St DESeq2	134.5	11.068	10	84.885	184.114	12.151	1.699e-05
Sr_St LimmaVoom - Hi_St edgeR	86.166	19.172	10	0.232	172.101	4.494	0.049
Sr_St LimmaVoom - Sr_St edgeR	96.5	11.068	10	46.885	146.114	8.718	0.000
Sr_St LimmaVoom - Hi_Ht LimmaVoom	-88.000	15.653	10	-158.165	-17.834	-5.621	0.011

Model 1 (Pipeline-Specific Filtered):

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_St DESeq2 - Hi_Ht DESeq2	-91.666	10.693	10	-139.598	-43.734	-8.572	0.000
K_K2 DESeq2 - Hi_St DESeq2	73.333	10.693	10	25.401	121.265	6.857	0.002
Sl_Sl2 DESeq2 - Hi_St DESeq2	57	10.693	10	9.067	104.932	5.330	0.016
Sr_Ht DESeq2 - Hi_St DESeq2	107.333	10.693	10	59.401	155.265	10.037	9.700e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_Ht DESeq2 - Sl_Sl2 DESeq2	50.333	10.693	10	2.401	98.265	4.706	0.037
Sr_St DESeq2 - Hi_Ht DESeq2	-71.000	10.693	10	-118.932	-23.067	-6.639	0.003
Sr_St DESeq2 - K_K2 DESeq2	-52.666	10.693	10	-100.598	-4.734	-4.925	0.027
Sr_St DESeq2 - Sr_Ht DESeq2	-86.666	10.693	10	-134.598	-38.734	-8.104	0.000
Hi_Ht edgeR - Hi_Ht DESeq2	54.666	7.561	10	20.773	88.559	7.229	0.001
Hi_Ht edgeR - Hi_St DESeq2	146.333	13.097	10	87.628	205.038	11.172	3.688e-05
Hi_Ht edgeR - K_K2 DESeq2	73.000	13.097	10	14.295	131.704	5.573	0.011
Hi_Ht edgeR - Sl_Sl2 DESeq2	89.333	13.097	10	30.628	148.038	6.820	0.002
Hi_Ht edgeR - Sr_St DESeq2	125.666	13.097	10	66.961	184.371	9.595	0.000
Hi_St edgeR - Hi_St DESeq2	54.666	7.561	10	20.773	88.559	7.229	0.001
Hi_St edgeR - Hi_Ht edgeR	-91.666	10.693	10	-139.598	-43.734	-8.572	0.000
K_K2 edgeR - Hi_St DESeq2	128	13.097	10	69.295	186.704	9.773	0.000
K_K2 edgeR - K_K2 DESeq2	54.666	7.561	10	20.773	88.559	7.229	0.001
K_K2 edgeR - Sl_Sl2 DESeq2	70.999	13.097	10	12.295	129.704	5.421	0.014
K_K2 edgeR - Sr_St DESeq2	107.333	13.097	10	48.628	166.038	8.195	0.000
K_K2 edgeR - Hi_St edgeR	73.333	10.693	10	25.401	121.265	6.857	0.002
Sl_Sl2 edgeR - Hi_St DESeq2	111.666	13.097	10	52.961	170.371	8.526	0.000

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sl_Sl2 edgeR - Sl_Sl2 DESeq2	54.666	7.561	10	20.773	88.559	7.229	0.001
Sl_Sl2 edgeR - Sr_St DESeq2	90.999	13.097	10	32.295	149.704	6.948	0.002
Sl_Sl2 edgeR - Hi_St edgeR	57	10.693	10	9.067	104.932	5.330	0.016
Sr_Ht edgeR - Hi_Ht DESeq2	70.333	13.097	10	11.628	129.038	5.370	0.015
Sr_Ht edgeR - Hi_St DESeq2	162	13.097	10	103.295	220.704	12.369	1.437e-05
Sr_Ht edgeR - K_K2 DESeq2	88.666	13.097	10	29.961	147.371	6.769	0.002
Sr_Ht edgeR - Sl_Sl2 DESeq2	105	13.097	10	46.295	163.704	8.017	0.000
Sr_Ht edgeR - Sr_Ht DESeq2	54.666	7.561	10	20.773	88.559	7.229	0.001
Sr_Ht edgeR - Sr_St DESeq2	141.333	13.097	10	82.628	200.038	10.791	5.057e-05
Sr_Ht edgeR - Hi_St edgeR	107.333	10.693	10	59.401	155.265	10.037	9.700e-05
Sr_Ht edgeR - Sl_Sl2 edgeR	50.333	10.693	10	2.401	98.265	4.706	0.037
Sr_St edgeR - Hi_St DESeq2	75.333	13.097	10	16.628	134.038	5.751	0.009
Sr_St edgeR - Sr_St DESeq2	54.666	7.561	10	20.773	88.559	7.229	0.001
Sr_St edgeR - Hi_Ht edgeR	-71.000	10.693	10	-118.932	-23.067	-6.639	0.003
Sr_St edgeR - K_K2 edgeR	-52.666	10.693	10	-100.598	-4.734	-4.925	0.027
Sr_St edgeR - Sr_Ht edgeR	-86.666	10.693	10	-134.598	-38.734	-8.104	0.000
Hi_Ht LimmaVoom - Hi_Ht DESeq2	172.333	7.561	10	138.440	206.226	22.790	3.765e-08

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_Ht LimmaVoom - Hi_St DESeq2	264	13.097	10	205.295	322.704	20.157	1.614e-07
Hi_Ht LimmaVoom - K_K2 DESeq2	190.666	13.097	10	131.961	249.371	14.557	2.975e-06
Hi_Ht LimmaVoom - Sl_Sl2 DESeq2	207	13.097	10	148.295	265.704	15.805	1.388e-06
Hi_Ht LimmaVoom - Sr_Ht DESeq2	156.666	13.097	10	97.961	215.371	11.961	1.967e-05
Hi_Ht LimmaVoom - Sr_St DESeq2	243.333	13.097	10	184.628	302.038	18.579	3.501e-07
Hi_Ht LimmaVoom - Hi_Ht edgeR	117.666	7.561	10	83.773	151.559	15.561	1.596e-06
Hi_Ht LimmaVoom - Hi_St edgeR	209.333	13.097	10	150.628	268.038	15.983	1.257e-06
Hi_Ht LimmaVoom - K_K2 edgeR	136	13.097	10	77.295	194.704	10.383	7.154e-05
Hi_Ht LimmaVoom - Sl_Sl2 edgeR	152.333	13.097	10	93.628	211.038	11.631	2.552e-05
Hi_Ht LimmaVoom - Sr_Ht edgeR	102	13.097	10	43.295	160.704	7.787	0.000
Hi_Ht LimmaVoom - Sr_St edgeR	188.666	13.097	10	129.961	247.371	14.405	3.294e-06
Hi_St LimmaVoom - Hi_Ht DESeq2	80.666	13.097	10	21.961	139.371	6.159	0.005
Hi_St LimmaVoom - Hi_St DESeq2	172.333	7.561	10	138.440	206.226	22.790	3.765e-08
Hi_St LimmaVoom - K_K2 DESeq2	99	13.097	10	40.295	157.704	7.558	0.001
Hi_St LimmaVoom - Sl_Sl2 DESeq2	115.333	13.097	10	56.628	174.038	8.806	0.000
Hi_St LimmaVoom - Sr_Ht DESeq2	64.999	13.097	10	6.295	123.704	4.962	0.026
Hi_St LimmaVoom - Sr_St DESeq2	151.666	13.097	10	92.961	210.371	11.580	2.657e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi_St LimmaVoom - Hi_St edgeR	117.666	7.561	10	83.773	151.559	15.561	1.596e-06
Hi_St LimmaVoom - Sl_Sl2 edgeR	60.666	13.097	10	1.961	119.371	4.632	0.040
Hi_St LimmaVoom - Sr_St edgeR	97	13.097	10	38.295	155.704	7.406	0.001
Hi_St LimmaVoom - Hi_Ht LimmaVoom	-91.666	10.693	10	-139.598	-43.734	-8.572	0.000
K_K2 LimmaVoom - Hi_Ht DESeq2	154	13.097	10	95.295	212.704	11.758	2.308e-05
K_K2 LimmaVoom - Hi_St DESeq2	245.666	13.097	10	186.961	304.371	18.757	3.216e-07
K_K2 LimmaVoom - K_K2 DESeq2	172.333	7.561	10	138.440	206.226	22.790	3.765e-08
K_K2 LimmaVoom - Sl_Sl2 DESeq2	188.666	13.097	10	129.961	247.371	14.405	3.294e-06
K_K2 LimmaVoom - Sr_Ht DESeq2	138.333	13.097	10	79.628	197.038	10.562	6.138e-05
K_K2 LimmaVoom - Sr_St DESeq2	225	13.097	10	166.295	283.704	17.179	6.809e-07
K_K2 LimmaVoom - Hi_Ht edgeR	99.333	13.097	10	40.628	158.038	7.584	0.001
K_K2 LimmaVoom - Hi_St edgeR	191	13.097	10	132.295	249.704	14.583	2.926e-06
K_K2 LimmaVoom - K_K2 edgeR	117.666	7.561	10	83.773	151.559	15.561	1.596e-06
K_K2 LimmaVoom - Sl_Sl2 edgeR	134	13.097	10	75.295	192.704	10.231	8.171e-05
K_K2 LimmaVoom - Sr_Ht edgeR	83.666	13.097	10	24.961	142.371	6.388	0.004
K_K2 LimmaVoom - Sr_St edgeR	170.333	13.097	10	111.628	229.038	13.005	8.900e-06
K_K2 LimmaVoom - Hi_St LimmaVoom	73.333	10.693	10	25.401	121.265	6.857	0.002

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sl_Sl2 LimmaVoom - Hi_Ht DESeq2	137.666	13.097	10	78.961	196.371	10.511	6.411e-05
Sl_Sl2 LimmaVoom - Hi_St DESeq2	229.333	13.097	10	170.628	288.038	17.510	5.807e-07
Sl_Sl2 LimmaVoom - K_K2 DESeq2	156	13.097	10	97.295	214.704	11.911	2.047e-05
Sl_Sl2 LimmaVoom - Sl_Sl2 DESeq2	172.333	7.561	10	138.440	206.226	22.790	3.765e-08
Sl_Sl2 LimmaVoom - Sr_Ht DESeq2	122	13.097	10	63.295	180.704	9.315	0.000
Sl_Sl2 LimmaVoom - Sr_St DESeq2	208.666	13.097	10	149.961	267.371	15.932	1.293e-06
Sl_Sl2 LimmaVoom - Hi_Ht edgeR	82.999	13.097	10	24.295	141.704	6.337	0.004
Sl_Sl2 LimmaVoom - Hi_St edgeR	174.666	13.097	10	115.961	233.371	13.336	6.977e-06
Sl_Sl2 LimmaVoom - K_K2 edgeR	101.333	13.097	10	42.628	160.038	7.737	0.000
Sl_Sl2 LimmaVoom - Sl_Sl2 edgeR	117.666	7.561	10	83.773	151.559	15.561	1.596e-06
Sl_Sl2 LimmaVoom - Sr_Ht edgeR	67.333	13.097	10	8.628	126.038	5.141	0.020
Sl_Sl2 LimmaVoom - Sr_St edgeR	154	13.097	10	95.295	212.704	11.758	2.308e-05
Sl_Sl2 LimmaVoom - Hi_St LimmaVoom	57	10.693	10	9.067	104.932	5.330	0.016
Sr_Ht LimmaVoom - Hi_Ht DESeq2	188	13.097	10	129.295	246.704	14.354	3.408e-06
Sr_Ht LimmaVoom - Hi_St DESeq2	279.666	13.097	10	220.961	338.371	21.353	8.557e-08
Sr_Ht LimmaVoom - K_K2 DESeq2	206.333	13.097	10	147.628	265.038	15.754	1.429e-06
Sr_Ht LimmaVoom - Sl_Sl2 DESeq2	222.666	13.097	10	163.961	281.371	17.001	7.428e-07

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_Ht LimmaVoom - Sr_Ht DESeq2	172.333	7.561	10	138.440	206.226	22.790	3.765e-08
Sr_Ht LimmaVoom - Sr_St DESeq2	259	13.097	10	200.295	317.704	19.775	1.958e-07
Sr_Ht LimmaVoom - Hi_Ht edgeR	133.333	13.097	10	74.628	192.038	10.180	8.544e-05
Sr_Ht LimmaVoom - Hi_St edgeR	225	13.097	10	166.295	283.704	17.179	6.809e-07
Sr_Ht LimmaVoom - K_K2 edgeR	151.666	13.097	10	92.961	210.371	11.580	2.657e-05
Sr_Ht LimmaVoom - Sl_Sl2 edgeR	168	13.097	10	109.295	226.704	12.827	1.016e-05
Sr_Ht LimmaVoom - Sr_Ht edgeR	117.666	7.561	10	83.773	151.559	15.561	1.596e-06
Sr_Ht LimmaVoom - Sr_St edgeR	204.333	13.097	10	145.628	263.038	15.601	1.559e-06
Sr_Ht LimmaVoom - Hi_St LimmaVoom	107.333	10.693	10	59.401	155.265	10.037	9.700e-05
Sr_Ht LimmaVoom - Sl_Sl2 LimmaVoom	50.333	10.693	10	2.401	98.265	4.706	0.037
Sr_St LimmaVoom - Hi_Ht DESeq2	101.333	13.097	10	42.628	160.038	7.737	0.000
Sr_St LimmaVoom - Hi_St DESeq2	193	13.097	10	134.295	251.704	14.736	2.649e-06
Sr_St LimmaVoom - K_K2 DESeq2	119.666	13.097	10	60.961	178.371	9.136	0.000
Sr_St LimmaVoom - Sl_Sl2 DESeq2	136	13.097	10	77.295	194.704	10.383	7.154e-05
Sr_St LimmaVoom - Sr_Ht DESeq2	85.666	13.097	10	26.961	144.371	6.540	0.003
Sr_St LimmaVoom - Sr_St DESeq2	172.333	7.561	10	138.440	206.226	22.790	3.765e-08
Sr_St LimmaVoom - Hi_St edgeR	138.333	13.097	10	79.628	197.038	10.562	6.138e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr_St LimmaVoom - K_K2 edgeR	65.000	13.097	10	6.295	123.704	4.962	0.026
Sr_St LimmaVoom - Sl_Sl2 edgeR	81.333	13.097	10	22.628	140.038	6.210	0.005
Sr_St LimmaVoom - Sr_St edgeR	117.666	7.561	10	83.773	151.559	15.561	1.596e-06
Sr_St LimmaVoom - Hi_Ht LimmaVoom	-71.000	10.693	10	-118.932	-23.067	-6.639	0.003
Sr_St LimmaVoom - K_K2 LimmaVoom	-52.666	10.693	10	-100.598	-4.734	-4.925	0.027
Sr_St LimmaVoom - Sr_Ht LimmaVoom	-86.666	10.693	10	-134.598	-38.734	-8.104	0.000

Table S2.2: Analysis of Variance (ANOVA) summaries of model 2 for each filtering method. Sum of Squares values were used to calculate percent variability explained by the predictors that are reported in the text.

	Sum Sq	Df	F value	Pr (>F)
Soft Filtered				
Aligner	784	1	10.02	0. 0.025
Counter	12352	1	157.79	0.000
DGE Program	51430	2	328.49	0.000
Counter:DGE	1401	2	8.95	0.022
Residuals	391	5		
Hard Filtered				
Aligner	616	1	13.2	0.015
Counter	16280	1	349.9	0.000
DGE Program	32651	2	350.8	0.000
Counter:DGE	2056	2	22.1	0.003
Residuals	233	5		
Pipeline-Specific Filtered				
Aligner	990.1	1	181	0.000
Counter	23852.1	1	4350	0.000
DGE Program	67086.5	2	6117	0.000
Counter:DGE	1372.2	2	125	0.000
Residuals	27.4	5		

Table S2.3: Linear model 2 summaries for each filtering method. Intercept is the grand mean. Estimates are mean differences in the number of significant DEGs from the grand mean or intercept averaged over all other predictors (i.e., Hisat2 estimate is the average number of significant DEGs across both counters and all 3 DGE programs when the aligner is Hisat2, minus the grand mean). R-squared values were rounded during export from R. Values below estimates in parentheses are standard error estimates.

	Soft	Hard	Pipeline
(Intercept)	1062.58 ***	956.33 ***	1042.25 ***
	(2.55)	(1.97)	(0.68)
Hisat2	-8.08 *	7.17 *	-9.08 ***
	(2.55)	(1.97)	(0.68)
HTSeq	32.08 ***	36.83 ***	44.58 ***
	(2.55)	(1.97)	(0.68)
DESeq2	-72.83 ***	-55.33 ***	-79.25 ***
	(3.61)	(2.78)	(0.96)
EdgeR	-13.08 *	-14.58 **	-21.00 ***
	(3.61)	(2.78)	(0.96)
HTSeq:DESeq2	9.67 *	1.17	13.92 ***
	(3.61)	(2.78)	(0.96)
HTSeq:EdgeR	-15.08 **	-16.58 **	-1.83
	(3.61)	(2.78)	(0.96)
N	12	12	12
R2	0.99	1.00	1.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table S2.4 Contrast between pipelines from emmeans for Model 2. Pairwise contrasts between each pipeline. Estimates are differences between predicted (estimated marginal) means for pipeline comparisons. Contrasts have been filtered for significant values less than or equal to 0.05 to reduce table sizes. SE – standard error. see Table 2.1 for program abbreviations.

Model 2 (Soft Filtered):

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi St DESeq2 - Hi Ht DESeq2	-83.500	8.847	5	-129.317	-37.682	-9.437	0.003
Hi St DESeq2 - Sr Ht DESeq2	-99.666	10.216	5	-152.572	-46.760	-9.755	0.003
Sr St DESeq2 - Hi Ht DESeq2	-67.333	10.216	5	-120.239	-14.427	-6.590	0.018
Sr St DESeq2 - Sr Ht DESeq2	-83.500	8.847	5	-129.317	-37.682	-9.437	0.003
Hi Ht edgeR - Hi St DESeq2	118.5	8.847	5	72.682	164.317	13.393	0.000
Hi Ht edgeR - Sr St DESeq2	102.333	10.216	5	49.427	155.239	10.016	0.002
Sr Ht edgeR - Hi St DESeq2	134.666	10.216	5	81.760	187.572	13.181	0.000
Sr Ht edgeR - Sr St DESeq2	118.5	8.847	5	72.682	164.317	13.393	0.000
Hi St edgeR - Hi St DESeq2	84.5	8.847	5	38.682	130.317	9.550	0.003
Hi St edgeR - Sr St DESeq2	68.333	10.216	5	15.427	121.239	6.688	0.017
Sr St edgeR - Hi St DESeq2	100.666	10.216	5	47.760	153.572	9.853	0.002

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr St edgeR - Sr St DESeq2	84.5	8.847	5	38.682	130.317	9.550	0.003
Hi Ht LimmaVoom - Hi Ht DESeq2	154.5	8.847	5	108.682	200.317	17.461	0.000
Hi Ht LimmaVoom - Sr Ht DESeq2	138.333	10.216	5	85.427	191.239	13.540	0.000
Hi Ht LimmaVoom - Hi St DESeq2	238	8.847	5	192.182	283.817	26.899	2.871e-05
Hi Ht LimmaVoom - Sr St DESeq2	221.833	10.216	5	168.927	274.739	21.713	5.148e-05
Hi Ht LimmaVoom - Hi Ht edgeR	119.5	8.847	5	73.682	165.317	13.506	0.000
Hi Ht LimmaVoom - Sr Ht edgeR	103.333	10.216	5	50.427	156.239	10.114	0.002
Hi Ht LimmaVoom - Hi St edgeR	153.5	8.847	5	107.682	199.317	17.348	0.000
Hi Ht LimmaVoom - Sr St edgeR	137.333	10.216	5	84.427	190.239	13.442	0.000
Sr Ht LimmaVoom - Hi Ht DESeq2	170.666	10.216	5	117.760	223.572	16.704	0.000
Sr Ht LimmaVoom - Sr Ht DESeq2	154.5	8.847	5	108.682	200.317	17.461	0.000
Sr Ht LimmaVoom - Hi St DESeq2	254.166	10.216	5	201.260	307.072	24.877	3.376e-05
Sr Ht LimmaVoom - Sr St DESeq2	238	8.847	5	192.182	283.817	26.899	2.871e-05
Sr Ht LimmaVoom - Hi Ht edgeR	135.666	10.216	5	82.760	188.572	13.279	0.000
Sr Ht LimmaVoom - Sr Ht edgeR	119.5	8.847	5	73.682	165.317	13.506	0.000
Sr Ht LimmaVoom - Hi St edgeR	169.666	10.216	5	116.760	222.572	16.607	0.000
Sr Ht LimmaVoom - Sr St edgeR	153.5	8.847	5	107.682	199.317	17.348	0.000

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi St LimmaVoom - Hi Ht DESeq2	79.499	8.847	5	33.682	125.317	8.985	0.004
Hi St LimmaVoom - Sr Ht DESeq2	63.333	10.216	5	10.427	116.239	6.199	0.023
Hi St LimmaVoom - Hi St DESeq2	163	8.847	5	117.182	208.817	18.422	0.000
Hi St LimmaVoom - Sr St DESeq2	146.833	10.216	5	93.927	199.739	14.372	0.000
Hi St LimmaVoom - Hi St edgeR	78.5	8.847	5	32.682	124.317	8.872	0.004
Hi St LimmaVoom - Sr St edgeR	62.333	10.216	5	9.427	115.239	6.101	0.025
Hi St LimmaVoom - Hi Ht LimmaVoom	-75	8.847	5	-120.817	-29.182	-8.476	0.005
Hi St LimmaVoom - Sr Ht LimmaVoom	-91.166	10.216	5	-144.072	-38.260	-8.923	0.004
Sr St LimmaVoom - Hi Ht DESeq2	95.666	10.216	5	42.760	148.572	9.363	0.003
Sr St LimmaVoom - Sr Ht DESeq2	79.499	8.847	5	33.682	125.317	8.985	0.004
Sr St LimmaVoom - Hi St DESeq2	179.166	10.216	5	126.260	232.072	17.536	0.000
Sr St LimmaVoom - Sr St DESeq2	163	8.847	5	117.182	208.817	18.422	0.000
Sr St LimmaVoom - Hi Ht edgeR	60.666	10.216	5	7.760	113.572	5.938	0.028
Sr St LimmaVoom - Hi St edgeR	94.666	10.216	5	41.760	147.572	9.266	0.003
Sr St LimmaVoom - Sr St edgeR	78.5	8.847	5	32.682	124.317	8.872	0.004
Sr St LimmaVoom - Hi Ht LimmaVoom	-58.833	10.216	5	-111.739	-5.927	-5.758	0.032
Sr St LimmaVoom - Sr Ht LimmaVoom	-75	8.847	5	-120.817	-29.182	-8.476	0.005

Model 2 (Hard Filtered):

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi St DESeq2 - Hi Ht DESeq2	-76.000	6.821	5	-111.324	-40.675	-11.141	0.001
Hi St DESeq2 - Sr Ht DESeq2	-61.666	7.876	5	-102.456	-20.876	-7.828	0.008
Sr St DESeq2 - Hi Ht DESeq2	-90.333	7.876	5	-131.123	-49.543	-11.468	0.001
Sr St DESeq2 - Sr Ht DESeq2	-76.000	6.821	5	-111.324	-40.675	-11.141	0.001
Hi Ht edgeR - Hi St DESeq2	99	6.821	5	63.675	134.324	14.512	0.000
Hi Ht edgeR - Sr St DESeq2	113.333	7.876	5	72.543	154.123	14.388	0.000
Sr Ht edgeR - Hi St DESeq2	84.666	7.876	5	43.876	125.456	10.748	0.001
Sr Ht edgeR - Sr St DESeq2	99	6.821	5	63.675	134.324	14.512	0.000
Hi St edgeR - Hi St DESeq2	58.5	6.821	5	23.175	93.824	8.575	0.005
Hi St edgeR - Sr St DESeq2	72.833	7.876	5	32.043	113.623	9.246	0.003
Hi St edgeR - Hi Ht edgeR	-40.5	6.821	5	-75.824	-5.175	-5.937	0.028
Sr St edgeR - Hi St DESeq2	44.166	7.876	5	3.376	84.956	5.607	0.036
Sr St edgeR - Sr St DESeq2	58.5	6.821	5	23.175	93.824	8.575	0.005

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr St edgeR - Hi Ht edgeR	-54.833	7.876	5	-95.623	-14.043	-6.961	0.014
Sr St edgeR - Sr Ht edgeR	-40.5	6.821	5	-75.824	-5.175	-5.937	0.028
Hi Ht LimmaVoom - Hi Ht DESeq2	139.5	6.821	5	104.175	174.824	20.449	6.911e-05
Hi Ht LimmaVoom - Sr Ht DESeq2	153.833	7.876	5	113.043	194.623	19.529	8.996e-05
Hi Ht LimmaVoom - Hi St DESeq2	215.5	6.821	5	180.175	250.824	31.591	1.883e-05
Hi Ht LimmaVoom - Sr St DESeq2	229.833	7.876	5	189.043	270.623	29.178	2.384e-05
Hi Ht LimmaVoom - Hi Ht edgeR	116.5	6.821	5	81.175	151.824	17.078	0.000
Hi Ht LimmaVoom - Sr Ht edgeR	130.833	7.876	5	90.043	171.623	16.609	0.000
Hi Ht LimmaVoom - Hi St edgeR	157	6.821	5	121.675	192.324	23.015	4.138e-05
Hi Ht LimmaVoom - Sr St edgeR	171.333	7.876	5	130.543	212.123	21.751	5.109e-05
Sr Ht LimmaVoom - Hi Ht DESeq2	125.166	7.876	5	84.376	165.956	15.890	0.000
Sr Ht LimmaVoom - Sr Ht DESeq2	139.5	6.821	5	104.175	174.824	20.449	6.911e-05
Sr Ht LimmaVoom - Hi St DESeq2	201.166	7.876	5	160.376	241.956	25.539	3.193e-05
Sr Ht LimmaVoom - Sr St DESeq2	215.5	6.821	5	180.175	250.824	31.591	1.883e-05
Sr Ht LimmaVoom - Hi Ht edgeR	102.166	7.876	5	61.376	142.956	12.970	0.000
Sr Ht LimmaVoom - Sr Ht edgeR	116.5	6.821	5	81.175	151.824	17.078	0.000
Sr Ht LimmaVoom - Hi St edgeR	142.666	7.876	5	101.876	183.456	18.112	0.000

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr Ht LimmaVoom - Sr St edgeR	157	6.821	5	121.675	192.324	23.015	4.138e-05
Hi St LimmaVoom - Sr Ht DESeq2	49.333	7.876	5	8.543	90.123	6.263	0.022
Hi St LimmaVoom - Hi St DESeq2	111	6.821	5	75.675	146.324	16.271	0.000
Hi St LimmaVoom - Sr St DESeq2	125.333	7.876	5	84.543	166.123	15.911	0.000
Hi St LimmaVoom - Hi St edgeR	52.5	6.821	5	17.175	87.824	7.696	0.009
Hi St LimmaVoom - Sr St edgeR	66.833	7.876	5	26.043	107.623	8.484	0.005
Hi St LimmaVoom - Hi Ht LimmaVoom	-104.5	6.821	5	-139.824	-69.175	-15.319	0.000
Hi St LimmaVoom - Sr Ht LimmaVoom	-90.166	7.876	5	-130.956	-49.376	-11.447	0.001
Sr St LimmaVoom - Hi St DESeq2	96.666	7.876	5	55.876	137.456	12.272	0.001
Sr St LimmaVoom - Sr St DESeq2	111	6.821	5	75.675	146.324	16.271	0.000
Sr St LimmaVoom - Sr St edgeR	52.5	6.821	5	17.175	87.824	7.696	0.009
Sr St LimmaVoom - Hi Ht LimmaVoom	-118.833	7.876	5	-159.623	-78.043	-15.086	0.000
Sr St LimmaVoom - Sr Ht LimmaVoom	-104.5	6.821	5	-139.824	-69.175	-15.319	0.000

Model 2 (Pipeline-Specific Filtered):

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Sr Ht DESeq2 - Hi Ht DESeq2	18.166	1.351	5	11.165	25.167	13.437	0.000
Hi St DESeq2 - Hi Ht DESeq2	-117.000	2.341	5	-129.126	-104.873	-49.964	1.074e-06
Hi St DESeq2 - Sr Ht DESeq2	-135.166	2.703	5	-149.168	-121.164	-49.989	1.070e-06
Sr St DESeq2 - Hi Ht DESeq2	-98.833	2.703	5	-112.835	-84.831	-36.552	9.966e-06
Sr St DESeq2 - Sr Ht DESeq2	-117.000	2.341	5	-129.126	-104.873	-49.964	1.074e-06
Sr St DESeq2 - Hi St DESeq2	18.166	1.351	5	11.165	25.167	13.437	0.000
Hi Ht edgeR - Hi Ht DESeq2	42.499	2.341	5	30.373	54.626	18.149	0.000
Hi Ht edgeR - Sr Ht DESeq2	24.333	2.703	5	10.331	38.335	8.999	0.004
Hi Ht edgeR - Hi St DESeq2	159.5	2.341	5	147.373	171.626	68.114	5.154e-07
Hi Ht edgeR - Sr St DESeq2	141.333	2.703	5	127.331	155.335	52.270	8.154e-07
Sr Ht edgeR - Hi Ht DESeq2	60.666	2.703	5	46.664	74.668	22.436	4.515e-05
Sr Ht edgeR - Sr Ht DESeq2	42.499	2.341	5	30.373	54.626	18.149	0.000
Sr Ht edgeR - Hi St DESeq2	177.666	2.703	5	163.664	191.668	65.707	5.179e-07
Sr Ht edgeR - Sr St DESeq2	159.5	2.341	5	147.373	171.626	68.114	5.154e-07
Sr Ht edgeR - Hi Ht edgeR	18.166	1.351	5	11.165	25.167	13.437	0.000
Hi St edgeR - Hi Ht DESeq2	-43.000	2.341	5	-55.126	-30.873	-18.363	0.000
Hi St edgeR - Sr Ht DESeq2	-61.166	2.703	5	-75.168	-47.164	-22.621	4.384e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi St edgeR - Hi St DESeq2	74	2.341	5	61.873	86.126	31.601	1.881e-05
Hi St edgeR - Sr St DESeq2	55.833	2.703	5	41.831	69.835	20.649	6.561e-05
Hi St edgeR - Hi Ht edgeR	-85.5	2.341	5	-97.626	-73.373	-36.512	1.002e-05
Hi St edgeR - Sr Ht edgeR	-103.666	2.703	5	-117.668	-89.664	-38.339	7.568e-06
Sr St edgeR - Hi Ht DESeq2	-24.833	2.703	5	-38.835	-10.831	-9.184	0.004
Sr St edgeR - Sr Ht DESeq2	-43.000	2.341	5	-55.126	-30.873	-18.363	0.000
Sr St edgeR - Hi St DESeq2	92.166	2.703	5	78.164	106.168	34.086	1.402e-05
Sr St edgeR - Sr St DESeq2	74	2.341	5	61.873	86.126	31.601	1.881e-05
Sr St edgeR - Hi Ht edgeR	-67.333	2.703	5	-81.335	-53.331	-24.902	3.369e-05
Sr St edgeR - Sr Ht edgeR	-85.5	2.341	5	-97.626	-73.373	-36.512	1.002e-05
Sr St edgeR - Hi St edgeR	18.166	1.351	5	11.165	25.167	13.437	0.000
Hi Ht LimmaVoom - Hi Ht DESeq2	153.499	2.341	5	141.373	165.626	65.552	5.181e-07
Hi Ht LimmaVoom - Sr Ht DESeq2	135.333	2.703	5	121.331	149.335	50.051	1.061e-06
Hi Ht LimmaVoom - Hi St DESeq2	270.5	2.341	5	258.373	282.626	115.516	5.137e-07
Hi Ht LimmaVoom - Sr St DESeq2	252.333	2.703	5	238.331	266.335	93.321	5.137e-07
Hi Ht LimmaVoom - Hi Ht edgeR	111	2.341	5	98.873	123.126	47.402	1.582e-06
Hi Ht LimmaVoom - Sr Ht edgeR	92.833	2.703	5	78.831	106.835	34.333	1.358e-05

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi Ht LimmaVoom - Hi St edgeR	196.5	2.341	5	184.373	208.626	83.915	5.137e-07
Hi Ht LimmaVoom - Sr St edgeR	178.333	2.703	5	164.331	192.335	65.953	5.175e-07
Sr Ht LimmaVoom - Hi Ht DESeq2	171.666	2.703	5	157.664	185.668	63.488	5.229e-07
Sr Ht LimmaVoom - Sr Ht DESeq2	153.499	2.341	5	141.373	165.626	65.552	5.181e-07
Sr Ht LimmaVoom - Hi St DESeq2	288.666	2.703	5	274.664	302.668	106.759	5.137e-07
Sr Ht LimmaVoom - Sr St DESeq2	270.5	2.341	5	258.373	282.626	115.516	5.137e-07
Sr Ht LimmaVoom - Hi Ht edgeR	129.166	2.703	5	115.164	143.168	47.770	1.490e-06
Sr Ht LimmaVoom - Sr Ht edgeR	111	2.341	5	98.873	123.126	47.402	1.582e-06
Sr Ht LimmaVoom - Hi St edgeR	214.666	2.703	5	200.664	228.668	79.391	5.137e-07
Sr Ht LimmaVoom - Sr St edgeR	196.5	2.341	5	184.373	208.626	83.915	5.137e-07
Sr Ht LimmaVoom - Hi Ht LimmaVoom	18.166	1.351	5	11.165	25.167	13.437	0.000
Hi St LimmaVoom - Hi Ht DESeq2	88.499	2.341	5	76.373	100.626	37.793	8.250e-06
Hi St LimmaVoom - Sr Ht DESeq2	70.333	2.703	5	56.331	84.335	26.011	3.075e-05
Hi St LimmaVoom - Hi St DESeq2	205.5	2.341	5	193.373	217.626	87.758	5.137e-07
Hi St LimmaVoom - Sr St DESeq2	187.333	2.703	5	173.331	201.335	69.282	5.148e-07
Hi St LimmaVoom - Hi Ht edgeR	46	2.341	5	33.873	58.126	19.644	8.690e-05
Hi St LimmaVoom - Sr Ht edgeR	27.833	2.703	5	13.831	41.835	10.293	0.002

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Hi St LimmaVoom - Hi St edgeR	131.5	2.341	5	119.373	143.626	56.156	6.118e-07
Hi St LimmaVoom - Sr St edgeR	113.333	2.703	5	99.331	127.335	41.914	4.134e-06
Hi St LimmaVoom - Hi Ht LimmaVoom	-65.000	2.341	5	-77.126	-52.873	-27.758	2.685e-05
Hi St LimmaVoom - Sr Ht LimmaVoom	-83.166	2.703	5	-97.168	-69.164	-30.757	2.054e-05
Sr St LimmaVoom - Hi Ht DESeq2	106.666	2.703	5	92.664	120.668	39.449	6.315e-06
Sr St LimmaVoom - Sr Ht DESeq2	88.499	2.341	5	76.373	100.626	37.793	8.250e-06
Sr St LimmaVoom - Hi St DESeq2	223.666	2.703	5	209.664	237.668	82.719	5.137e-07
Sr St LimmaVoom - Sr St DESeq2	205.5	2.341	5	193.373	217.626	87.758	5.137e-07
Sr St LimmaVoom - Hi Ht edgeR	64.166	2.703	5	50.164	78.168	23.731	3.784e-05
Sr St LimmaVoom - Sr Ht edgeR	46	2.341	5	33.873	58.126	19.644	8.690e-05
Sr St LimmaVoom - Hi St edgeR	149.666	2.703	5	135.664	163.668	55.352	6.385e-07
Sr St LimmaVoom - Sr St edgeR	131.5	2.341	5	119.373	143.626	56.156	6.118e-07
Sr St LimmaVoom - Hi Ht LimmaVoom	-46.833	2.703	5	-60.835	-32.831	-17.320	0.000
Sr St LimmaVoom - Sr Ht LimmaVoom	-65.000	2.341	5	-77.126	-52.873	-27.758	2.685e-05
Sr St LimmaVoom - Hi St LimmaVoom	18.166	1.351	5	11.165	25.167	13.437	0.000

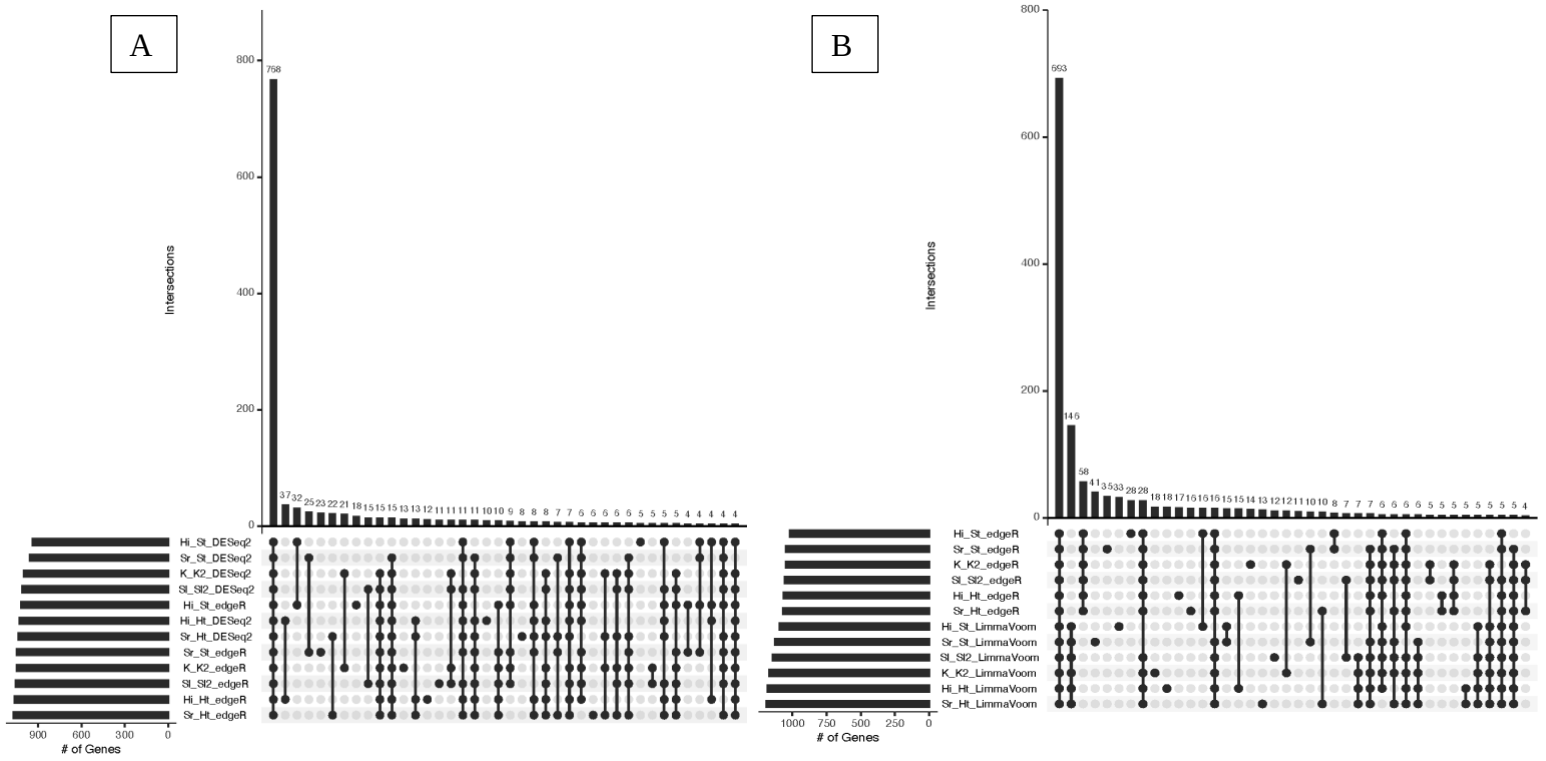

Figure S2.2. High overlap between DGE programs, particularly those using Negative Binomial Models

Upset plots of DEGs for soft filtering. DEG intersections for pipelines including (A) EdgeR or DESeq2 and (B) EdgeR or Limma-Voom. Horizontal bars represent the total number of DEGs for a pipeline. Filled circles indicate the specific pipelines being intersected and the vertical bars are the number of DEGs in that intersection.

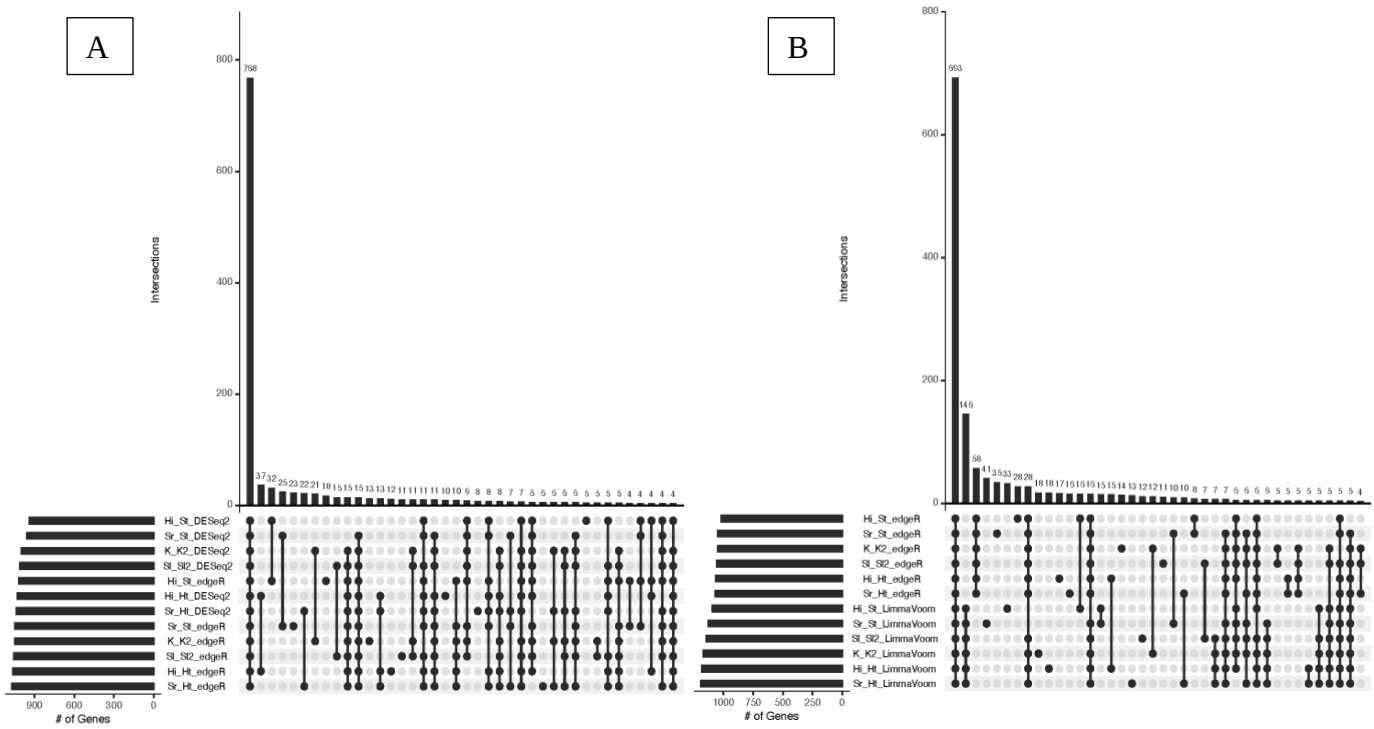

Figure S2.3. High overlap between DGE programs, particularly those using Negative Binomial Models

Upset plots of DEGs for hard filtering. DEG intersections for pipelines including (A) EdgeR or DESeq2 and (B) EdgeR or Limma-Voom. Horizontal bars represent the total number of DEGs for a pipeline. Filled circles indicate the specific pipelines being intersected and the vertical bars are the number of DEGs in that intersection.

Chapter 3

Table S3.1: Details of provenance and demographics of Anoles established using these methods. Reared category indicates whether an individual was caught in the wild or born in the lab. Y – year, M – month, D – Day are organism ages at the point of establishment. AD = reproductively mature adults, sampled at end of breeding season.

Species	Sex	Origin	Reared	Age
sagrei	M	Bahamas	Lab	3Y
sagrei	M	Bahamas	Lab	3Y
sagrei	F	Bahamas	Lab	3Y
sagrei	M	Bahamas	Lab	3Y
sagrei	F	Bahamas	Lab	3Y
sagrei	M	Bahamas	Lab	3Y
sagrei	F	Bahamas	Lab	3Y
sagrei	F	Bahamas	Lab	3Y
sagrei	F	Bahamas	Lab	3Y
sagrei	F	Bahamas	Lab	2Y
sagrei	M	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	F	Bahamas	Lab	4Y
sagrei	F	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	F	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	F	Bahamas	Lab	4Y
sagrei	F	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	4Y
sagrei	F	Bahamas	Lab	4Y
sagrei	M	Bahamas	Lab	1.5Y
sagrei	F	Bahamas	Lab	1.5Y
sagrei	M	Bahamas	Lab	1.5Y

sagrei	F	Bahamas	Lab	1.5Y
sagrei	F	Bahamas	Lab	1.5Y
sagrei	F	Bahamas	Lab	1.5Y
sagrei	M	Bahamas	Lab	1.5Y
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	1.5Y
sagrei	M	Bahamas	Lab	1.5Y
sagrei	F	Bahamas	Lab	1.5Y
sagrei	F	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	M	Bahamas	Lab	30D
sagrei	F	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M

sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	5Y
sagrei	M	Bahamas	Lab	5Y
sagrei	M	Bahamas	Lab	5Y
sagrei	M	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	5Y
sagrei	M	Bahamas	Lab	7M
sagrei	M	Bahamas	Lab	7M
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	5Y
sagrei	F	Bahamas	Lab	5Y
carolinensis	M	Coral Gables, FL	Wild	AD
carolinensis	M	Coral Gables, FL	Wild	AD
carolinensis	M	Coral Gables, FL	Wild	AD
carolinensis	F	Coral Gables, FL	Wild	AD
carolinensis	F	Coral Gables, FL	Wild	AD
carolinensis	F	Coral Gables, FL	Wild	AD
carolinensis	F	Coral Gables, FL	Wild	AD
carolinensis	M	Coral Gables, FL	Wild	AD
equestris	F	Coral Gables, FL	Wild	AD
equestris	F	Coral Gables, FL	Wild	AD
equestris	F	Coral Gables, FL	Wild	AD
equestris	M	Coral Gables, FL	Wild	AD

equestris	F	Coral Gables, FL	Wild	AD
equestris	M	Coral Gables, FL	Wild	AD
equestris	M	Coral Gables, FL	Wild	AD
cybotes	M	Port Mayaca, FL	Wild	AD
cybotes	F	Port Mayaca, FL	Wild	AD
cybotes	M	Port Mayaca, FL	Wild	AD
cybotes	F	Port Mayaca, FL	Wild	AD
cybotes	F	Port Mayaca, FL	Wild	AD
cybotes	F	Port Mayaca, FL	Wild	AD
cybotes	M	Port Mayaca, FL	Wild	AD
cybotes	M	Port Mayaca, FL	Wild	AD
chlorocyanus	F	Parkland, FL	Wild	AD
chlorocyanus	M	Parkland, FL	Wild	AD
chlorocyanus	M	Parkland, FL	Wild	AD
chlorocyanus	M	Parkland, FL	Wild	AD
chlorocyanus	M	Parkland, FL	Wild	AD
chlorocyanus	F	Parkland, FL	Wild	AD
distichus	F	Coral Gables, FL	Wild	AD
distichus	F	Coral Gables, FL	Wild	AD
distichus	F	Coral Gables, FL	Wild	AD
distichus	F	Coral Gables, FL	Wild	AD

APPENDIX 2: Abstracts of Additional Contributions

Sparkman, A.M., A.D. Clark, L.J. Brummett, et al. "Convergence in reduced body size, head size, and blood glucose in three island reptiles," Ecology & Evolution, 2018;8:6169–6182. https://doi.org/10.1002/ece3.4171

Many oceanic islands harbor diverse species that differ markedly from their mainland relatives with respect to morphology, behavior, and physiology. A particularly common morphological change exhibited by a wide range of species on islands worldwide involves either a reduction in body size, termed island dwarfism, or an increase in body size, termed island gigantism. While numerous instances of dwarfism and gigantism have been well documented, documentation of other morphological changes on islands remains limited. Furthermore, we lack a basic understanding of the physiological mechanisms that underlie these changes, and whether they are convergent. A major hypothesis for the repeated evolution of dwarfism posits selection for smaller, more efficient body sizes in the context of low resource availability. Under this hypothesis, we would expect the physiological mechanisms known to be downregulated in model organisms exhibiting small body sizes due to dietary restriction or artificial selection would also be downregulated in wild species exhibiting dwarfism on islands. We measured body size, relative head size, and circulating blood glucose in three species of reptiles—two snakes and one lizard—in the California Channel Islands relative to mainland populations. Collating data from 6 years of study, we found that relative to mainland population the island populations had smaller body size (i.e., island dwarfism), smaller head sizes relative to body size, and lower levels of blood glucose, although with

some variation by sex and year. These findings suggest that the island populations of these three species have independently evolved convergent physiological changes (lower glucose set point) corresponding to convergent changes in morphology that are consistent with a scenario of reduced resource availability and/or changes in prey size on the islands. This provides a powerful system to further investigate ecological, physiological, and genetic variables to elucidate the mechanisms underlying convergent changes in life history on islands.

Westfall, A.K., R.S. Telemeco, M.B. Grizante, D.S. Waits, A.D. Clark, et al. "A chromosome-level genome assembly for the Eastern fence lizard (Sceloporus undulatus), a reptile model for physiological and evolutionary ecology," GigaScience. 2021; https://doi.org/10.1093/gigascience/giab066

High-quality genomic resources facilitate investigations into behavioral ecology, morphological and physiological adaptations, and the evolution of genomic architecture. Lizards in the genus Sceloporus have a long history as important ecological, evolutionary, and physiological models, making them a valuable target for the development of genomic resources. We present a high-quality chromosome-level reference genome assembly, SceUnd1.0 (using 10X Genomics Chromium, HiC, and Pacific Biosciences data), and tissue/developmental stage transcriptomes for the eastern fence lizard, Sceloporus undulatus. We performed synteny analysis with other snake and lizard assemblies to identify broad patterns of chromosome evolution including the fusion of micro- and macrochromosomes. We also used this new assembly to provide improved

reference-based genome assemblies for 34 additional Sceloporus species. Finally, we used RNAseq and whole-genome resequencing data to compare 3 assemblies, each representing an increased level of cost and effort: Supernova Assembly with data from 10X Genomics Chromium, HiRise Assembly that added data from HiC, and PBJelly Assembly that added data from Pacific Biosciences sequencing. We found that the Supernova Assembly contained the full genome and was a suitable reference for RNAseq and single-nucleotide polymorphism calling, but the chromosome-level scaffolds provided by the addition of HiC data allowed synteny and whole-genome association mapping analyses. The subsequent addition of PacBio data doubled the contig N50 but provided negligible gains in scaffold length. These new genomic resources provide valuable tools for advanced molecular analysis of an organism that has become a model in physiology and evolutionary ecology.

Clark, A.D. & L.S. Stevison. "Learning R for Biologists: a mini course grab-bag for instructors," CourseSource, Accepted – In revisions May 2022

As biology becomes more data driven, teaching students data literacy skills has become central to biology programs. Despite a wealth of online resources that teach researchers how to use R, there are few that offer practical laboratory-based exercises, with teaching resources such as keys, learning objectives, and assessment materials. Here, we present a modular set of lessons and lab activities to help teach R through the platform of R Studio. Both softwares are free and open source making this curriculum highly accessible across various institutions. This curriculum was developed over several years of teaching a graduate level computational biology course. In response to the pandemic, the class was shifted to be completely online. These resources were

then migrated to GitHub to make them broadly accessible to anyone wanting to learn R for the analysis of biological datasets. In the following year, these resources were used to teach the course in a flipped format, which is the lesson plan presented here. In general, students responded well to the flipped format, which used class time to conduct live coding demos and work through challenges with the instructor and teaching assistant. Overall, students were able to use these skills to analyze and interpret data, as well as produce publication quality graphics. While the modules presented range from very basic, doing simples summary statistics and plotting, to quite advanced, where R is integrated onto the command line, teachers should feel free to pick and choose which elements to incorporate into their own curriculum.